Correlating Refugee Border Crossings with Internet Search Data

Panos Kostakos, Abhinay Pandya, Mourad Oussalah, Simo Hosio, Arash Sattari
Center for Ubiquitous Computing
University of Oulu
Oulu, Finland
panos.kostakos;bhinay.pandya;
mourad.oussalah;
simo.hosio@oulu.fi

Vassilis Kostakos, Niels van Berkel, Christoph Breidbach
School of Computing and Information Systems
University of Melbourne
Melbourne, Australia
vassilis.kostakos;christoph.breidbac
@unimelb.edu.au;n.vanberkel@stu
dent.unimelb.edu.au

Olga Kyriakouli
Department of Informatics and Telematics
Harokopio University of Athens
Athens, Greece
itp16401@hua.gr

***Abstract*—Can Internet search data be used as a proxy to predict refugee mobility? The soaring refugee death toll in Europe creates an urgent need for novel tools that monitor and forecast refugee flows. This study investigates the correlation between refugee mobility data and Internet search data from Google Trends. Google Trends is a freely accessible tool that provides access to Internet search data by analyzing a sample of all web queries. In our study, we surveyed refugees in Greece (entry point) and in Finland (destination point) to identify what search queries they had used during their travel. Next, we conducted time series analysis on Google search data to investigate whether interest in user-defined search queries correlated with the levels of refugee arrival data recorded by the United Nations High Commissioner for Refugees (UNHCR). Results indicate that the reuse of internet search data considerably improves the predictive power of the models.**

Google Trends; Internet search data, Refugees, Greece, Turkey.

I. INTRODUCTION

In 2016 over 3,000 refugees and migrants lost their lives crossing the Mediterranean Sea entering in Europe [1]. Monitoring migration movements plays a key role in effectively managing migration and refugee waves from war zones and enabling emergency response teams and volunteers to better prepare for such crises. Surveilling refugee and migration movement requires the gathering of information from local communities, the media, and humanitarian organizations stationed in war zones [2]. Much of this data collection effort depends on individual cases reported by local communities and mass movements between borders. However, over the past years, the increasing demand for irregular travel from conflict zones has fueled a substantial illegal market that facilities clandestine movement, movement that often goes under the radar of humanitarian organizations and local organizations. As a result, the movement of both migrants and refugees is volatile, lacks predictability, and in combination with poor infrastructure in transit countries, contributes to human lose and evolving humanitarian crises.

Recent research has shown that online behavior, such as queries conducted using the Google search engine, are correlated with geopolitical activities and refuge mobility [3]. Building on the intuition that online digital footprints can be used as proxy for capturing offline social dynamics [4], we explore how internet search volume can be used to enhance forecasting models of patterns of migrant and refugee movement into Europe. We surveyed refugees to identify keywords and topics they searched during their travels. Then, we collected data about the popularity of these keywords through the Google Trends application. Finally, we developed five models to predict refugee arrivals in Greece between August 2015 and November 2017[1].

II. MATERIALS AND METHODS

Predicting arrival times can drastically improve the local communities' ability to deal with humanitarian crises and early response teams to save lives. Among several indicators with the potential capability to predict arrival times and rates, Internet search data is both easily accessible and sufficiently dynamic. Monitoring other factors such as change in economic conditions, the instigation of political coups, the development epidemics, etc. require too much effort and understanding of their respective domains. However, when the focus is only on predicting the time and rate of refugee arrival and not understanding the reasons and forces for migration, building a predictive model based on Internet search data is both efficient and accurate.

The goal of the predictive model is to take the input of internet search patterns in a given region and output an estimate of the arrival dates of refugees to an adjoining/nearby

[1] The code and data used in the experiments are available here: https://github.com/PanosKostakos/GTrends_refugees/

978-1-5386-2659-7/18/$31.00 ©2018 IEEE
DOI 10.1109/IRI.2018.00048

region. Essentially, in the context of machine learning, this predictive model learns a mapping between the internet search patterns and arrival dates. The parameters of this mapping or the unknown function are calculated from the historical information on arrivals and their corresponding internet search patterns. Since, mathematically speaking, a space of such functions is of infinite dimensions, machine learning researchers or data analyst experts assume a form of this function and use the appropriate algorithm to estimate its parameters.

TABLE I. LIST OF KEYWORDS SEARCHED BY RESPONDENTS WHILE TRAVELING FROM DESTINATION TO TARGET COUNTRIES.

Country	Language	Search Queries
Afghanistan	Farsi	Refugees, metro, bus, weather bus, weather Greece weather maps, google maps, roads buses Greece weather, Greece route Greece map.
Greece	Farsi	Metro weather bus trip, ship trip, Victoria park, go abroad weather, routes, Greek racism, Greece weather, weather, immigrants, BBC, google translate.
Iran	Farsi	Refugees camps, Europe route, weather, refugees number, routes refugees camps, Turkey, Greece, Victoria park, Greece weather, Greek camps, services, camps, google maps, weather, Facebook, bus Google
Turkey	Farsi	Turkish language, buses, google earth, weather, Aegean Sea, suburbs, clean camp, healthy camp, train tickets, ship tickets, turkey map, turkey weather, google maps, country population, culture, weather Aegean islands, sea condition weather.
Greece	Arabic	Life, refugees situations, Immigration jobs Hungary borders Macedonian borders, trains Transportation Greece routes immigrants Greece Mytilene Athens passage situation, passage safety Hungary route Finland.
Iraq	Arabic	Bus ticket, Europe borders, way Europe, way Finland buses, Europe transportation, google maps Europe immigration, route Europe, Europe borders, trip fee Finland life, Finland work, trains, buses, taxi, learn Finnish google maps, facebook weather, sea conditions, transportation, borders Greece posting buses, weather, sea crossing Europe directions turkey to Greece weather, sea conditions, route safety Europe rout, Europe immigration, Europe living, illegal immigration Railway stations, buses, humanitarian aid Weather, distance, buses, transportations Facebook, Whatsapp, Tango, Viber, Skype, imo, Google, maps, youtube, education and entertainment, Coco, Nature, science, tourism.
Fyrom	Arabic	Transportation routes, cross country.
Serbia	Arabic	Hungarian borders, transportation, cross Hungary Transportation Weather, Transportation Serbia, safe border, Facebook, Tango, Viber, Whatsapp, WeChat, Coco, imo, Skype, Google, Youtube, maps, translation apps, education and entertainment, Google Play.
Turkey	Arabic	Hotels, bus tickets Turkey, Greece hotels, trains, tickets prices Facebook Greece news, Macedonia news, Hungary news Sea conditions, Turkey Greece Transportation cross Greece cross sea, transportation, safety Finland, Europe way Greece route, weather, sea conditions Turkey weather, Turkey currency, Turkey Greece railway stations Greece weather, Greek Nature, Freedom, work, education, children education, family reunion, Athens route.

For the remainder of this section, we present the data sources and statistical methods used to predict the arrival of refugees and migrants in Greece. We are using data from the following sources:

Google Trend keywords. Google represents the largest market share in internet searches. Google Trends is a free online service provided by Google that offers normalized time series data for given keywords. Google Trends data capture search interest over time in given geographical locations. We collected weekly time series data for a set of keywords between August 2015 and November 2017. The keywords were identified using questionnaires.

Questionnaires. In order to get a more accurate understanding of what keywords refugees and migrants use before entering Europe, we conducted two small-scale surveys in Greece (Athens N=90) and Finland (Oulu N =50). The survey enabled us to understand smartphone usage and to define keywords and topics searched through these devises.

Border crossing hotspots. We have used location-based data to determine migration flows, entry points, and transit routes. Annotated maps from the International Organization for Migration [5] provide a fairly accurate overview of the main hotspots used for entering EU nations for the past few years. This resource enabled us to define which locations would be considered when extracting internet search data from Google Trends.

Daily refugee arrival times. We used the UN Refugee Agency website to retrieve daily arrival data from Greece and Italy [6]. Time series data are publicly available for the years 2015-2017. To build the prediction models, we converted this dataset into weekly data that can be easily fused with the data provided by Google Trends.

A. Selection of search queries

The selection of keywords to include and/or exclude from the analysis is a challenging task, and prior research has used different methods to overcome this problem [7]. In order to get accurate results, we ran a user study to elicit online behavioral patterns of refugees and migrants. A questionnaire was distributed to 140 refugees in two hotspots in Greece and Finland in order to determine how they have used technology when traveling to Europe. Below, we list the questions used for eliciting topics and keywords searched by the participants:

Q1. While in your hometown and before you started your journey to Europe, did you search the Internet for information specifically for this journey?

Q2. If YES, can you please list some of the exact keywords you entered into the online search engine?

Q3. Having left your hometown, what is the FIRST country you visited in your journey to Europe?

Q4. If you searched for information on the Internet while visiting the country above, can you please list some of the exact keywords you entered into the online search engine?

Based on their responses, we narrowed the queries in 6 different countries. Table 1 shows that most of the keywords can be organised into the following broader categories or topics: Transportation, weather, demographics, location, and social media. In this paper, we consider only the keywords used in Turkey.

B. Selection of geographical locations

Google Trends enables users to collect time series data of search interest from specific locations. This geographical granularity allows researchers to develop better insight into hyperlocal patters and offline social behavior. As shown in Figure 1, we used data from the International Organization for Migration (IOM) to identify the known migration routes and entry hotspots. This helped us to narrow down the geographical scope of our keyword search [5]. We identified the following eight interesting locations in Turkey (Google Trends code is provided in parenthesis): Edirne (TR-22); Canakkale Province (TR-17); Balikesir (TR-10); Izmir (TR-35); Aydin (TR-09); Mugla (TR-48); Alanya (TR-07); Mersin (TR-33).

Figure 1. Migration entry points and transit routes

C. Refugee and migrant arrival times

We conducted retrospective analysis of the refugee's and migrant's arrival patterns in Greece and Italy between August 2015 and November 2017 in order to determine variables that enabled us to predict mobility patterns. Arrival data are recorded by local authorities, and data are shared via the UNHCR Operational Portal. Because Google Trends data are organized into weekly intervals, we decided to transform the daily arrival data into weekly time series and rescale them to assume values between 0-100. In Figure 2 we visually inspect both time series and note the downward trend in the Greek dataset.

D. Machine learning and statistical methods

A number of tools were used for modelling the refugee arrival patterns in Greece. First, to determine what variables are strongly related with the arrival data in Greece we used simple Person correlation. We computed pairwise correlations between the arrivals in Greece and the keywords used by migrants or refugees in Turkey (extracted from the questionnaires). Second, to control for seasonal patterns we run the same correlation analysis, but this time using the arrival data from Italy. Third, we developed and evaluated the accuracy of the following five models for predicting arrival times in Greece: baseline model, multiple log linear regression, fully-grown decision tree, pruned decision tree, and random forest.

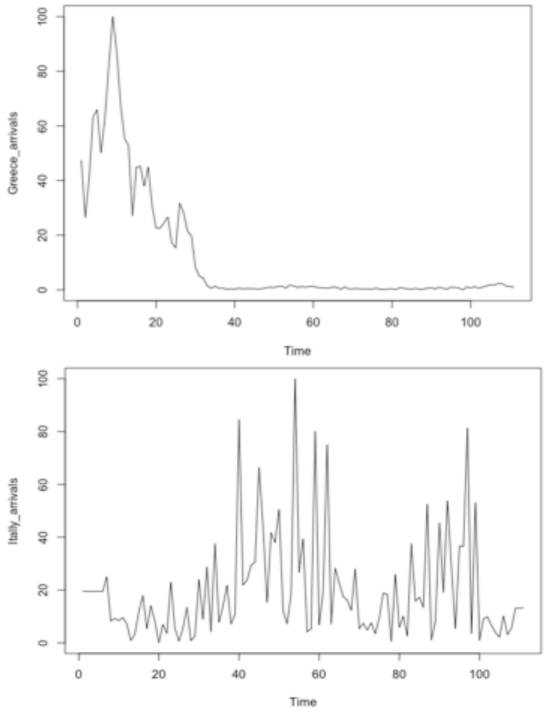

Figure 2. Transformed and rescaled refugee arrival data for Greece and Italy (Aug 2015- Nov 2017)

We study arrival times as a continues outcome rather than as binary (arrived/not-arrived). Hence, we experiment in the regression setting instead of classification. We performed a 70:30 random split of the 120 weeks and used 70% of the data for training models and 30% for evaluation. For evaluation, we are using RMSE (root mean squared error) and MAE (mean absolute error) as error metrics and observe how different set of keywords improve the power of the models. [8].

III. EXPERIMENTAL SETTING

Migration and refugee mobility depends on various geological, seasonal and spatiotemporal factors. This section outlines the experiments we have conducted to determine the best model for benchmarking and then predicting the continues outcome (arrival dates of migrants and refugees in Greece) using a set of predictors (Google search queries).

A. Selecting the independent variables

In order to develop effective models that use past Google search queries to improve on our prediction of arrival data, we first established the relationship between the dependent and

independent variables. Figure 3 shows the results of the correlation analysis using Pearson's *r* to determine what search terms correlate best with the arrival data in Greece and Italy. We note that the Greek arrivals correlate with the search queries, but arrivals in Italy do not have the linear relationship with the selected queries. We proceeded by compiling a list of 46 queries with r>0.5 that would be used for building and testing our models.

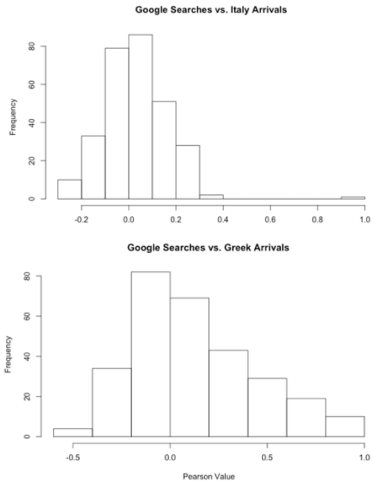

Figure 3. Distribution of correlated independent variables with the outcome variables in Greece and Italy

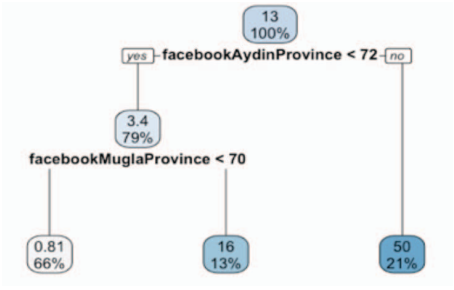

Figure 4. Pruned tree

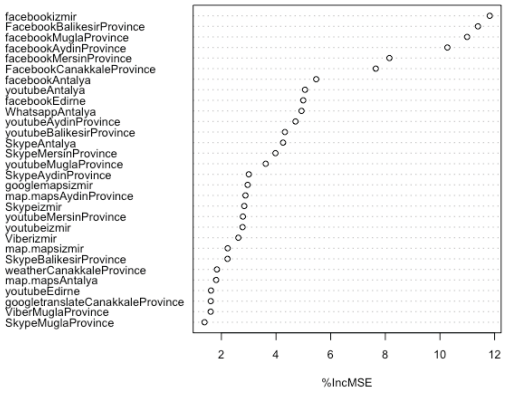

Figure 5. Variable importance plot

B. *Evaluation metrics*

The main metric for evaluating our models against the dependent variable will be the root mean squared error (RMSE) that gives more weight to larger disparities between the predicted and the observed values. This metric is sensitive to outliers, as larger errors have a disproportionately large effect on RMSE. As show in Figure 2 and Figure 3, we are modeling data that contain outliers. For a secondary metric, we will use the MAE (mean absolute error). This metric gives equal weight to the residuals, meaning that failure to predict 1000 arrivals is actually twice as bad as 5000 arrivals. The main advantage to this method is that it allows researchers to interpret the results based on the magnitude of the error when the trained model is fitted on unseen (test) data.

C. *Modeling*

We used the training data to build a baseline prediction model and subsequently fitted the model to the remaining test data. Given that refugee mobility is unregulated and uncontrolled, the baseline model was built on the intuition that the weekly mean arrivals in Greece is the best approximation for predicting future data (educated guess). Fitted on the test data, the baseline model returns RMSE= 22.41 and MAE=17.06.

Subsequently, we derived a multiple log linear regression model to enable prediction of weekly refugee arrival values from data gleaned from the 46 online search engine queries and observed the computed errors. We transformed the data using the ln(x+1) function in R to account for the fact that the dependant variable has zero values (weeks with no arrivals). The coefficient of multiple determination (R-squared) is 0.95, meaning that 95% of the variance in the arrival data can be explained by the set of predictors in the model. The adjusted R-squared (0.89) is notably lower than the original R-squared, indicating that some variables in the model are non-significant. This result was anticipated, given the large number of search queries. The most significant predictors were keywords associated with Facebook (Facebook in Mersin Province, p=0.00230; Facebook in the Mugla Province p= 0.02726) and YouTube (Youtube in Antalya Province, p=0.00133). Given the coefficient of Facebook search in the Mersin Province (-0.1453808), we reason that a unit increase in Google searches for Facebook in the Mersin region yields a decrease in the predicted number of refugees arriving in Greece by 13.5%. Compared to the educated guess of the baseline model, the regression model produced improved outcomes of RMSE=16.90 and MAE=7.00.

Next, we built a decision tree in R using a ten-fold cross-validation. In each iteration, 90% of the data is used for training and 10% for testing the model. We then prune the tree for the optimal number of splits. Figure 4 depicts that the tree was reduced to two splits and two variables relevant for predicting weekly arrivals in Greece. Facebook is again an important predictor that enables us to reason that if the weekly search popularity of Facebook in Aydin exceeds 72 units (as measured by the Google Trends scale), then the model predicts a high refugee arrival rate of 50 units. Both decision trees models produce improved error scores of RMSE=9.81 and MAE=3.29.

Lastly, to tackle our regression problem, we implemented a random forest with 655 trees. The final predicted values of the model derive from computing the weighted average of the value predicted by each tree. The accuracy of the model has improved significantly, with values of RMSE=7.42 and MAE=2.26 respectively. In Figure 5, we present the importance plot, observing that numerous Facebook search queries were found to be strong predictors of refugee arrivals in Greece. In the absence of these variables from the model, the error increases by a range of 10 to 40 percent.

Table 2 compares the models tested above: the random forest has achieved the best performance, and all the other models performed better that the baseline. In Figure 6, we illustrate the predicted and actual values detected by each model. The models are doing well in predicting about three units of the outcome variable, but fail to predict values beyond that threshold (green line). Irregular migration corresponds to various political, economic and environmental factors, and as discussed earlier in the paper, border crossing from Turkey to Greece has decreased substantially in the past few years, making it challenging for the models to learn this pattern. Despite the skewed distribution of the dependent outcome, the random forest model achieves greater accuracy. The overall results indicate that data of online search interests improve our prediction of refugee arrivals in Greece.

TABLE II. OVERALL ACCURACY OF THE MODELS

Method	RMSE	MAE
Baseline	22.41	17.06
Linear Regression	16.90	7.00
Full tree	9.82	3.30
Pruned tree	9.82	3.30
Random forest	7.43	2.27

IV. CONCLUSIONS

In this paper, we examined the question of whether Google search queries can be used to improve forecasting of refugee arrivals in Greece. Forecast models can offer authorities a window of opportunity to better deploy search and rescue operations. We tested four popular models against a baseline and evaluated the results on unseen data using two popular error metrics. While the models require additional parameterisation, preliminary conclusions show that social media queries improve the accuracy of the models.

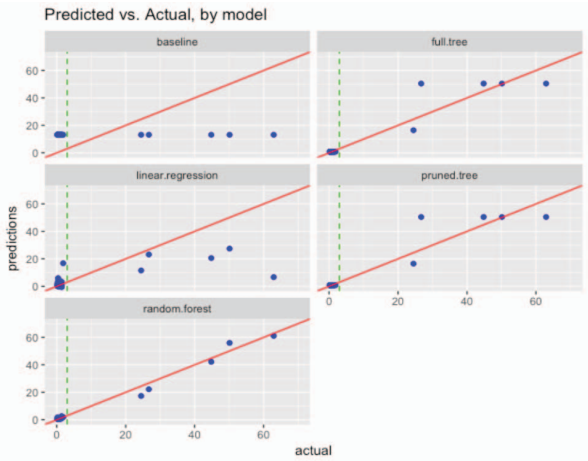

Figure 6. Predicted and actual values for all the models

ACKNOWLEDGMENT

This work is (partially) funded by the European Commission grant 770469-CUTLER and 645706-GRAGE

V. REFERENCES

[1] UNHCR Operational Portal, Mediterranean Situation, url: http://data2.unhcr.org/en/situations/mediterranean.

[2] P. Meier, "New information technologies and their impact on the humanitarian sector", *International Review of the Red Cross*, 2011, Volume 93, Issue 884, pp. 1239-1263.

[3] P. Connor., "The Digital Footprint of Europe's Refugees", Pew Research Center, 2017, url:http://www.pewglobal.org/2017/06/08/digital-footprint-of-europes-refugees.

[4] V. Kostakos, T. Juntunen, J. Goncalves, S. Hosio, and T. Ojala "Where Am I? Location Archetype Keyword Extraction from Urban Mobility Patterns", *Plos One*, 2013, url:http://journals.plos.org/plosone/article?id=10.1371/journal.pone.0063980.

[5] International Organization for Migration (IOM), url: http://migration.iom.int/europe.

[6] UNHCR Operational Portal, Mediterranean Situation, url: http://data2.unhcr.org/en/situations/mediterranean.

[7] S. V. Nuti, B. Wayda, I. Ranasinghe, S. Wang, R. P. Dreyer, S. I. Chen and K. Murugiah, "The Use of Google Trends in Health Care Research: A Systematic Review", *Plos One*, 2014 url: http://journals.plos.org/plosone/article?id=10.1371/journal.pone.0109583.

[8] C. J. Willmott and K. Matsuura (2005) 'Advantages of the mean absolute error (MAE) over the root mean square error (RMSE) in assessing average model performance', *Climate Research, 30(1): 79-82*