Search for a New Home: Refugee Mobility and Google Search

A. Ebru Sanliturk[1] and Francesco Billari[1]

[1]Bocconi University, Italy

Based on the assumption that trends of online search queries may indicate intentions and help to predict human behavior, this study addresses the general issue of analyzing, *nowcasting* and predicting migration decisions. We aim to contribute to the issue of settlement and re-settlement decisions across provinces using the case of urban refugees in Turkey as the case study and Google Trends data as the main estimator. The selected case study allows us to exploit the difference in the alphabet used by Turkish and Syrian citizens [1] as the method of differentiation between locals and Syrian refugees under the temporary protection status (hereafter 'under TP'). Using the alphabetical difference in the search for province names in Turkey and a unique dataset for the empirical analysis, the study seeks to answer two main research questions. The first and more general question is whether there is an association between the online search behavior and the settlement and re-settlement decisions of Syrians under TP. The second question is whether there is an association between the timing of the online search and the observed settlement / re-settlement.

Conceptual Framework

Online search trends and query frequency have attracted the attention of economists and social scientists in the last decade, as a possible estimator to predict future tendencies, i.e. *forecasting*, and events as well as to interpret the present, defined as *now-casting*. In economics, online search data is used as an estimator to forecast certain macroeconomic indicators, such as unemployment rate by online job searches (Ettredge, Gerdes, & Karuga, 2005), (Askitas & Zimmermann, 2009), inflation rate by expected inflation (Guzman, 2011) and for economic activity by search frequency (Choi & Varian, 2012). In the field of epidemiology, online search data is suggested as a measure to predict the present and/or immediate future. Deriving from the assumption that increased online search frequency for the early symptoms of contagious diseases indicate a risk of an outbreak, epidemiological research used online search data to now-cast infectious disease outbreaks such as influenza (Ginsberg et al., 2009), chickenpox (Pelat, Turbelin, Bar-Hen, Flahault, & Valleron, 2009) and salmonella (Brownstein, Freifeld, & Madoff, 2009).

More recently, demography literature also witnessed an increase in the use of online search trends data to analyze various types of demographic behavior. Internet search queries are found to be correlated with suicides (McCarthy, 2010), (Song, Song, An, Hayman, & Woo, 2014), (Solano et al., 2016) and also abortions (Reis & Brownstein, 2010) due to the need for information in case of restricted access. Google search trends are furthermore found to have a potential to forecast also fertility behavior (Billari, D'Amuri, & Marcucci, 2016).

Migration studies appears as a prominent field that benefited from the increased use of online data. This new source of data offered new insights where traditional data and official records were not sufficient. Especially the geo-location information provided by social media platforms such as Facebook (Zagheni, Weber, Gummadi, et al., 2017), Twitter (Hawelka et al., 2014) and LinkedIn (State, Rodriguez, Helbing, and Zagheni, 2014) became an important proxy to monitor and interpret migration flows.

[1]Latin alphabet is used for Turkish and Arabic alphabet is used for Arabic languages, spoken by the local and refugee communities respectively.

In migration studies, three studies stand out in terms of their relevance to the topic of this research; tracking mobility and migration through Google Trends data. Lin, Cranshaw, and Counts (2019) demonstrate a consistent and high correlation between the domestic migration predictions obtained through the analysis of search queries on Bing.com and official domestic migration records. Wladyka (2013), considers Google search queries related to migration to Spain as an indicator of intention for resettlement as well as a *forecast* measure for migration and uses online search data as a predictor of migration flows from Argentina, Colombia and Peru to Spain. Last but not least, Connor (2017) examines the relationship between queries made on Google by Syrian refugees and their movements from Syria to their destination country. He shows that a significant increase in the Google search frequency for "Greece" in Arabic in Turkey (where the predominant language is Turkish) is followed by an increase in the number of Syrian asylum seekers arriving in Greece and same pattern is observed in other EU countries for the search query "Germany" in Arabic.

Connor (2017)'s analysis on the case of Syrian refugees shows a sharp decrease in online search queries made with the intention to migrate to the EU following the introduction of EU-Turkey Deal in 2016. Thus, starting from the end point of this study and exploiting the difference in the alphabet rather than the language, we analyze whether a similar pattern exists for Syrians refugees' settlement and re-settlement decisions within Turkey. Thus, we compare the search queries for the name of every single province in Turkey (81 in total), both with Turkish and Arabic letters and both in Syria and Turkey with the official records on Syrian refugees under TP.

Background of the Case Study

The breakout of civil war in Syria occurred in 2011 and the situation escalated, with the consequent refugee influx into Mediterranean countries in 2014-2015. Since 2011, Turkey applied an open-door policy to Syrian refugees. As the urgency of the crisis became evident, the Turkish government signed the controversial readmission agreement with the EU (known as the EU-Turkey Deal) in 2015. The EU-Turkey Deal imposed Turkey to intercept refugees who wish to enter the EU borders through Turkey as well as strengthen its capabilities to register, accommodate and facilitate Syrian refugees. The EU-Turkey Deal is important for this research for two reasons; first it marks the starting point of the period analyzed in this research because as the Deal came into effect, Syrians under TP lost their passageway to the EU and had to remain in Turkey. Second, it also marks the date for the provincial-level data on Syrians under TP became publicly available.

Understanding the settlement/re-settlement patterns of Syrians under TP is important in the case of Turkey, as it provides a natural experiment. Syrians under TP constitute a great urban refugee community as only less than 5% of the total Syrians under TP in Turkey live in camp environments, while the remaining majority lives in urban areas. Syrians outside the camps are free to move inside the country, under the TP status.

Data and Methods

The period analyzed in this research, begins with the enactment of the EU-Turkey Deal in January 2016 ends on December 31, 2018, to allow for year fixed effects to be used for controls. The data used in this study relies on two main sources. The first source is Google Trends, providing normalized data for specified query index (name of provinces in Turkey with Latin and Arabic letters), in Syria and Turkey for the specified period. The alphabetical difference we exploit in this case study is important here, when the search query is the name of the province as these names remain by majority the same in different languages. Thus, the difference in the alphabet allows for a more accurate differentiation than the difference of language in this case. The second source of data for this study is the Directorate General of Migration Management (DGMM) that provides the official data on the number of refugees across provinces in Turkey. Per requirements of the EU-Turkey Deal, data on Syrian refugees is publicly available and continuously updated. However, the frequency of updates by DGMM is not standard and with each update, data on the preceding update disappears on the website. Believing that the available annual data may fail to capture the migration patterns in this case, we used the Wayback Machine that archives webpages through web crawls, to retrieve the data lost for public access (Arora, Li, Youtie, & Shapira, 2016). We further looked for the data on Syrians under TP across provinces in national/local media outlets and theses/dissertations published in Turkey and included the data to are dataset if the

proper DGMM citation and information on the week of the update is present. Using the sources above, we created a unique panel data set covering 3 years, which includes 157 weeks' Google search data in four categories (Arabic characters in Syria, Latin characters in Syria, Arabic characters in Turkey and Latin characters in Turkey) and 85 updates for the official number of refugees in 81 provinces. On this data set we first calculated the main explanatory variable as the proportion of search queries made in Turkey using Arabic letters over search queries using Latin letters (search query variable) to control for local and seasonal effect. Then we calculated the 4 lagged variables with one-week intervals and all control variables of search queries made in Syria together with their lagged variables with one-week intervals over the 157 weeks' period. Once all the search query variables were set, we matched the date of 85 updates of official registries with the data set of search query variables extending over 157 weeks. Last we dropped the weeks that we could not match with the data on Syrian refugees under protection. This way we ensured that time-lags are not lost or mismatched due to the shrinking sample size.

To test our hypotheses that (1) online search data frequency is correlated with the change in the number of Syrians under TP in a given province and (2) timing of the online search frequency is important to observe a change in the registries on Syrians under TP, we use an autoregressive fixed effects model. The reason for this choice relies on considering that in cases of nowcasting and forecasting the dependent variable is largely explained by its value at time (*t-1*) and considering the advantages of fixed effects model to absorb the effect of features that vary at province-level and year-level and allow a better interpretation of the explanatory variables.

$$ln(Number\ of\ Syrians\ u.\ TP)_{i[tj]} = \beta_1 \frac{Search\ query\ in\ ArabicA_{i[tj]}}{Search\ query\ in\ LatinA_{i[tj]}} + \alpha_i + \delta_i + e_{i[tj]} \quad (1)$$

The dependent variable in this baseline model is the log-transformation number of Syrians under TP in province *i*, at time *tj*, and the explanatory variable is the search query proportion at time *tj*, which denotes the Google searches made in the same week, i.e. official data on the number of refugees matched with the same week of Google search query data. In other words, this model tests for the *nowcasting* power of seach query variable. The province fixed effects are denoted by α_i and the year fixed effects are denoted by δ_i.

Next, in order to observe the potential effect of lags (*forecasting*), that is the difference between the time of the Google search and the time refugee-mobility appears on official data, we expand our baseline model to add the lagged variables. The lagged variables are denoted as *tj*-1, *tj*-2, *tj*-3 and *tj*-4. In our data, each lag refers to a one-week interval and we add a total of four lagged variables to the model, which at that stage becomes as follows;

$$\begin{aligned} ln(Number\ of\ Syrians\ u.\ TP)_{i[tj]} = {} & \beta_1 \frac{Search\ query\ in\ ArabicA_{i[tj]}}{Search\ query\ in\ LatinA_{i[tj]}} + \\ & \beta_2 \frac{Search\ query\ in\ ArabicA_{i[tj\text{-}1]}}{Search\ query\ in\ LatinA_{i[tj\text{-}1]}} + \beta_3 \frac{Search\ query\ in\ ArabicA_{i[tj\text{-}2]}}{Search\ query\ in\ LatinA_{i[tj\text{-}2]}} + \\ & \beta_4 \frac{Search\ query\ in\ ArabicA_{i[tj\text{-}3]}}{Search\ query\ in\ LatinA_{i[tj\text{-}3]}} + \beta_5 \frac{Search\ query\ in\ ArabicA_{i[tj\text{-}4]}}{Search\ query\ in\ LatinA_{i[tj\text{-}4]}} + \alpha_i + \delta_i + e_{it} \end{aligned} \quad (2)$$

Preliminary Results and Expected Outcomes

The results of our preliminary analysis, that is without introducing any control variables, confirm the two hypotheses and indicate a significant and positive correlation between the online search frequency and number of registered Syrians under TP as well as indicate that the timing of the online search matters for predicting mobility. The preliminary results are shown in the table below, where results are robust and standard errors are clustered at province level.

Variables		(1)		(2)
ln(Refugee#) (*t-1*)	0.967***	0.965***	0.966***	0.965***
	(0.00430)	(0.00721)	(0.00432)	(0.00718)
ArabicA/LatinA Query Proportion	-0.000941	-0.000932	-0.00185	-0.00179
	(0.00283)	(0.00289)	(0.00322)	(0.00325)
ArabicA/LatinA Query Proportion (*tj-1*)			0.00173*	0.00166*
			(0.000998)	(0.000996)
ArabicA/LatinA Query Proportion (*tj-2*)			0.00163	0.00175
			(0.00169)	(0.00172)
ArabicA/LatinA Query Proportion (*tj-3*)			0.000498	0.000606
			(0.00187)	(0.00180)
ArabicA/LatinA Query Proportion (*tj-4*)			0.00558***	0.00566***
			(0.00102)	(0.00112)
Year = 2017		0.00366*		0.00337*
		(0.00185)		(0.00184)
Year = 2018		0.00172		0.000830
		(0.00423)		(0.00420)
Constant	0.280***	0.289***	0.283***	0.286***
Observations	6,804	6,804	6,804	6,804
R-squared	0.961	0.961	0.961	0.961
Number of Provinces	81	81	81	81
Province fixed-effects	YES	YES	YES	YES
Year fixed-effects	NO	YES	NO	YES

Table 1: Preliminary Results [2]

The first column of the table, shows the results for the Arabic/Latin search query proportion (all searches made in Turkey) alone with province-level and without year fixed effects and in the column 2 year fixed effects are added. As observed in the first two columns, the simultaneous search query proportion variable is insignificant and the analysis shows no ground for nowcasting in this case. However, as the lagged variables are added in columns 3-4, we see that forecasting appears as a better estimation method and the search query proportion variable is positively significant for searches made one-week and four-weeks prior of the observed change in the official registries of Syrians under TP. The preliminary results suggest that one unit increase in the one-week prior search query proportion variable accounts for a 0.17% increase in the log-transformed number of Syrians under TP. In the case of one-month time lag where the observed association is strongest, the search query proportion variable accounts for a 0.57% increase in the log-transformed number of Syrians under TP.

As part of the initial robustness checks, we tried controlling for and replicating the model for the online searches made in Syria using Arabic and Latin letters and province names in Turkey as search query. Two important aspects emerged from these initial robustness checks. First, the search query variable is observed to be insignificant if the search is made in Syria, whether using Arabic or Latin letters but absence of online search in Syria as a dummy control is significant and negatively associated with our dependent variable. These results hint that a four-week lag may not be sufficient to observe the effect in the official registries if the search is made in Syria. Second, the positive and significant correlation between the number of Syrians under TP and online searches one-week (*tj-1*) and one-month prior in Turkey (*tj-4*) persists and is highest and strongest for the online searches made one-month prior in Turkey.

As the paper is in progress, currently we are continuing with the robustness checks. Our aim is to first expand our dataset further as possible to include more updates from the official registries. Second we aim to add more controls and further test the robustness of the preliminary results, possibly to include the effect of location and to cross-check the Google Trends data, the source of our main explanatory variables, with Google AdWords data. Last but not least, we aim to improve the interpretation of these results, together with possible policy implications.

References

Ettredge, M., Gerdes, J., & Karuga, G. (2005). Using web-based search data to predict macroeconomic statistics. *Communications of the ACM*, *48*(11), 87–92.

Askitas, N., & Zimmermann, K. F. (2009). Google econometrics and unemployment forecasting.

Brownstein, J. S., Freifeld, C. C., & Madoff, L. C. (2009). Digital disease detection—harnessing the Web for public health surveillance. *New England Journal of Medicine*, *360*(21), 2153–2157.

Ginsberg, J., Mohebbi, M. H., Patel, R. S., Brammer, L., Smolinski, M. S., & Brilliant, L. (2009). Detecting influenza epidemics using search engine query data. *Nature*, *457*(7232), 1012.

Pelat, C., Turbelin, C., Bar-Hen, A., Flahault, A., & Valleron, A.-J. (2009). More diseases tracked by using Google Trends. *Emerging infectious diseases*, *15*(8), 1327.

McCarthy, M. J. (2010). Internet monitoring of suicide risk in the population. *Journal of affective disorders*, *122*(3), 277–279.

Reis, B. Y., & Brownstein, J. S. (2010). Measuring the impact of health policies using Internet search patterns: the case of abortion. *BMC public health*, *10*(1), 514.

Guzman, G. (2011). Internet search behavior as an economic forecasting tool: The case of inflation expectations. *Journal of economic and social measurement*, *36*(3), 119–167.

Choi, H., & Varian, H. (2012). Predicting the present with Google Trends. *Economic Record*, *88*, 2–9.

Wladyka, D. (2013). *The queries to Google Search as predictors of migration flows from Latin America to Spain* (Doctoral dissertation).

Hawelka, B., Sitko, I., Beinat, E., Sobolevsky, S., Kazakopoulos, P., & Ratti, C. (2014). Geo-located Twitter as proxy for global mobility patterns. *Cartography and Geographic Information Science*, *41*(3), 260–271.

Song, T. M., Song, J., An, J.-Y., Hayman, L. L., & Woo, J.-M. (2014). Psychological and social factors affecting Internet searches on suicide in Korea: a big data analysis of Google search trends. *Yonsei medical journal*, *55*(1), 254–263.

State, B., Rodriguez, M., Helbing, D., & Zagheni, E. (2014). Migration of Professionals to the US. Springer Chamn.

Arora, S. K., Li, Y., Youtie, J., & Shapira, P. (2016). Using the wayback machine to mine websites in the social sciences: a methodological resource. *Journal of the Association for Information Science and Technology*, *67*(8), 1904–1915.

Billari, F., D'Amuri, F., & Marcucci, J. (2016). Forecasting births using Google. In *CARMA 2016: 1st International Conference on Advanced Research Methods in Analytics* (pp. 119–119). Editorial Universitat Politècnica de València.

Solano, P., Ustulin, M., Pizzorno, E., Vichi, M., Pompili, M., Serafini, G., & Amore, M. (2016). A Google-based approach for monitoring suicide risk. *Psychiatry research*, *246*, 581–586.

Connor, P. (2017). *The Digital Footprint of Europe's Refugees: Online Searches in 2015 and 2016 Open Window Into Path, Timing of Migrant Flows from Middle East to Europe*. Pew Research Center.

Zagheni, E., Weber, I., Gummadi, K., et al. (2017). Leveraging Facebook's advertising platform to monitor stocks of migrants. *Population and Development Review*, *43*(4), 721–734.

Lin, A. Y., Cranshaw, J., & Counts, S. (2019). Forecasting US Domestic Migration Using Internet Search Queries. In *The World Wide Web Conference* (pp. 1061–1072). ACM.