Journal of Population and Social Studies, October 2017 | vol. 25 no. 4: 312 - 327
DOI: 10.25133/JPSSv25n4.002

Queries to Google Search as Predictors of Migration Flows from Latin America to Spain

Dawid K. Wladyka[1]

This study evaluates the relationship between the changes in proportion of migration-related queries reported by Google Trends and changes in volume of migration flows between origin and destination countries. The study assesses if cost-free Google Trends improves the prediction of international migratory flows, and whether it could be proposed as a tool for organizations and policymakers. Previous research has used the activity of email users and other online services to track human mobility. At the same time, IP geolocation linked to Google Search has proven to be efficient in geographically tracking outbreaks of illnesses, as well as predicting changes in economic indicators and travel patterns. This research draws from both experiences. It uses a regression analysis of time series data to compare the popularity of migration related queries introduced to Google Search in Colombia, Argentina and Peru, to changes in a quantity of residents' registrations in Spain, performed by immigrants proceeding from these countries between the years 2005 and 2010. The results show a significant correlation and weak to moderate predictability for the lags of several months depending on the particular country. The findings demonstrate that trends in queries to Google Search provided by Google Trends might constitute a useful predictor of migration flows. At the same time, it indicates the need for further technological developments to improve analytical capacities.

Keywords *migration flows; IP geolocation; Google trends; Latin America; Spain*

Challenges in Estimations of International Migration Flows

The quality of statistical data on international migratory flows has been a constant challenge not only for researchers, but also for policymakers that have been eager to predict the number and origin of immigrants that seek new, at least temporary homes in their country, region or city. Despite the attempts to normalize international flows' statistics between countries the data is still unreliable and difficult to compare across borders (Kupiszewska & Nowok, 2008; Moses, 2012; United Nations [UN], 1998; Zagheni & Weber, 2012).

Data on migration flows are often generated a posteriori by comparison on time series of migration stocks as extracted from census data. The comparison of statistics based on definition of immigrant (or migration process) is often biased because of the use of different definitions by particular national statistical authorities. Another problematic issue is a time lag in which migration statistics are delivered (State, Weber & Zagheni, 2013). In countries where statistical authorities build their detailed data only on population censuses, carried out in a certain threshold of time (usually in 10-year periods), the availability of data makes them

[1] Sociology and Anthropology Department, University of Texas Rio Grande Valley, United States.
Email: dawid.wladyka@utrgv.edu

virtually useless to everyday analysis and policymaking. In some cases (e.g., developing countries), data on migration flows can be entirely unavailable because of financial limitations and/or lack of infrastructure. During the economic recession, financial cutbacks in the statistical authorities of developed countries could have resulted in a deterioration of the population statistics' quality.

On the other hand, already existent systems of capturing migration flows have a number of drawbacks inherent in the nature of migratory processes, as previously explained. These issues are primarily related to the undocumented immigrants' hesitation (or inability) to register because of status-related reasons. However, the delays between the displacement itself and registration apply also to the general immigrants' population. These individuals need time to familiarize themselves with appropriate procedures and pros/cons of fulfilling administrative formalities. Additionally, the economic calculations (i.e., international taxation rules) would negatively impact the decision of deregistering from the population censuses of countries of origin. Therefore, the flow estimations based on the population registers of origin countries also could provide misleading data (Zagheni & Weber, 2012).

Those shortcomings not only prevent local authorities from adjusting their policies in order to meet the social reality produced by an immediate future of immigration flows, but they also result in a lack of awareness about the current demographic trends of the region or city in question. This lack of data makes it practically impossible to adjust to the demand for use of local services (such as schools, medical facilities) or policies related to housing, welfare state and employment (UN, 1998). Furthermore, it would not only be useful to have a raw number or estimated proportion of immigrants, but also their geographic origins, as it could indicate cultural traits useful from the point of view of receiving country authorities. This information would allow undertaking such basic challenges as to train and employ an appropriate number of relevant translators and social-cultural mediators. Additionally, this data is valid information for receiving, as well as sending countries since the migrant's remittances directly influence the economy of origin, especially during economic recessions (UN, 1998).

Luckily, the last two decades, along with a growing number of international migration flows, brought about the development and dissemination of information technology. Therefore, the spatial mobility of individuals has already begun to be analyzed with the use of Internet Protocol (IP) geolocation technology. The same technology is used by Google Search engines in order to geographically track incoming queries, which can be further examined by the Google Trends application. Therefore, this research examines if, at the current stage of the technological development and data accessibility, the latter mentioned tool can be a useful addition to other methods of estimating and predicting international migration flows.

Geolocation in Human Mobility Tracking

The notion of geolocation bears multiple meanings depending on the particular discipline and service it describes. In this study, geolocation (also called geotagging) refers to technological means (e.g., IP address, cellular network, Wi-Fi network, Global Positioning System [GPS] device coordinates) that allow researchers to determine an individual's location, and to track the change of this location at various geographical and administrative levels (e.g., international, cities, blocks or streets). The focus here is on IP address geolocation which is one of the most sophisticated ways, but at the same time common and quick, to determine the location of the Internet service's user (King, 2010).

During these early endeavors with geolocation, scholars' efforts have been principally focused on the description of short distance displacement patterns at the city and regional level. The well-known study carried out by Ferrari, Rosi, Mamei and Zambonelli (2011) used Twitter data in order to shed light on New York's urban displacement patterns. Also, mobile phone data and applications, such as Google Latitude and Foursquare, that allow sharing of location data with friends have been subsequently used for description of short distance mobility displacements. On the other hand, the attention of tourism oriented researchers has been drawn to analysis of spatial data. Photos and online posts which included geo-references in web-based services like Flickr and CouchSurfing, together with mobile network data have provided new tools for the development tourism industry (State et al., 2013).

Geolocation data on displacements, "long" with both distance and duration, have begun to be tracked and analyzed, too. Statistical information provided by websites like **wheresgeorge.com**, which tracks one-dollar bills, and **geocaching.com**, which tracks various items (e.g., books, teddy bears) worldwide, have been used to examine displacement patterns (Zagheni & Weber, 2012). However, recently more complex databases have been used to track and analyze human mobility on the basis of geolocation of IP addresses. These databases are numerical names of every particular device connected to the Internet. The IP number usually corresponds to the geographic region of the Internet provider and can point not only to the country, but also to smaller units like regions and cities. Although geolocation by the IP number has some shortcomings (e.g., use of proxy servers that are located in different physical places), they have proved to be a reliable source of spatial data, especially at a country level (Gueye, Uhlig & Fdida 2007; Hui et al., 2010).

Svantesson (2004) and, building on his work, King (2010) provided accessible and straightforward descriptions of this process, formulating the following explanation. In its most basic form, the user-side process of browsing the Internet is composed of either introducing the website address called Uniform Resource Locator (URL) into the address bar of the Internet browser, or clicking a hyperlink containing the information about a desired URL. After completing either of these actions, the browser sends an access-request to the server hosting the desired content in order to provide it to the user. While there are some in-between technical steps that contain the mechanical translation of the human accessible URL to the numeric IP address by the appropriate Domain Name Service (DNS), these will be omitted here since they are not essential to the studied feature. After receiving the access-request, the server hosting the content makes a location-request for comparison of the user's IP address with the database of the provider of the geolocation service. The provider of the geolocation service returns an educated guess regarding the user's geographic location, and sends it to the server hosting desired content.

It should also be noted that the IP address geolocation process described above could be substituted with other forms of detecting a user's location. The websites often take advantage of the user's browser settings regarding country, time zone and language of the user. Still, this kind of location guess is less accurate than the one based on the IP address location which reaches 99% accuracy at the country level, and more than 90% accuracy for cities worldwide (King, 2010; Svantesson, 2004).

Google Search, Google Trends and Prediction of Real Events

The conceptual foundation of many studies cited herein has been derived from the study dedicated to detecting the epidemics of influenza by using Google Search queries (Ginsberg et al., 2009). This study developed a state-level model of estimation of influenza-like-illness

(ILI) outbreaks in the United States through the use of a five-year database of hundreds of billions of individual searches through the Google Search engine. It is important to mention that Google Search is a search engine responsible for more than 78% of Internet searches in the U.S. (FairSearch, 2013), about 90% in Latin America (ComScore, 2011) and 93% in Europe (FairSearch, 2013).

While IP geolocation was the technological foundation of the study, Google users that searched for the influenza-related terms were not expected to possess an account on a Google website. Although the disadvantage is that no knowledge about those who searched for given terms could be acquired, the simple aim of constructing a model that would instantly warn about the geographical location of an influenza outbreak could be achieved based on searches, which is virtually impossible to achieve through any other research method. The study — rooted in the assumption that a person with symptoms of ILI would start to seek out information regarding drugs and treatment methods — provided the model that is able to reliably predict the outbreaks of influenza long before national health authorities.

The successful predictive role of the queries in this study is underscored by the further development and improvement of Google Trends free online application that provides a frequency with which the given query has been entered in the Google Search in relation to the total search volume in the geographic location of the query origin. Successful predictions of outbreaks of other illnesses like chickenpox or salmonella have been based on data collected through Google Trends (Mohebbi et al., 2011). Other economy-related predictions also have been successful, specifically those focused on the retail sales of various goods. Likewise, and coming from a similar perspective as this study, the relation between Google searches and unemployment claims also appears relevant (Choi & Varian, 2009b; Mohebbi et al., 2011).

The most promising data regarding IP geolocation and travel displacements is based on research related to the common use of Internet and Google Search for travel planning. Analyzing Hong Kong visitors' data, Choi and Varian (2009b) found that Google Trends could predict visits to tourist destinations. There are three crucial findings in the Choi and Varian (2009b) study that support the assumptions underlying this research. First, the successful prediction of travel data by Choi and Varian (2009b) paves the way to the assumption that Google Search can be useful in predicting migration flows. Second, the query search time series has been lagged which adds the predictive power to the analysis of travel to Hong Kong. Third, the study on Hong Kong visitors uses data directly from Google Trends' open access, web-based application instead of the raw search data from the Google database, like in the case of the pioneering influenza study (Ginsberg et al., 2009). This supports the assumption that the combination of appropriately selected IP geolocated Google queries could provide an evaluation of interest in migration from particular location to particular destination.

Estimation of Migration Flows with Yahoo! Accounts IP Geolocation

Following some initial successes of IP address geolocation in previously mentioned fields, Zagheni and Weber (2012) attempted to estimate international migration rates with the help of geolocated email messages, using a large sample of IP geolocated messages sent between 2009 and 2011 from Yahoo! email accounts. The country from which most of the email messages were sent was considered the user's country of residence. This data was compared to national statistics on immigration rates gathered in 11 European countries. According to authors' findings, their predictions of age and gender were consistent with the archival data. Zagheni and Weber (2012) therefore suggest that geolocation of email messages can be an added value to worldwide migration statistics.

Instead of tracking where the email messages originated, another relevant study examined the locations where Yahoo! users log in during a one-year period. In particular, State et al. (2013) analyzed global country-to-country flows and attempted to illustrate current human flows worldwide. A number of important findings have been underlined in this study, among them the similarity in patterns of short and long-term displacements and relatively strong connections between some former colonies and metropolises, like Spain or England.

Whether the findings of the above mentioned studies do indeed support and open new perspectives for IP geolocation as a tool in migration flow research, they are not without disadvantages. One of the primary concerns is directly related to the data source used, which is the Yahoo! database. While a set of services related and branded as Yahoo! is a popular service in the United States, it is not as common in other parts of the world. This may provide a handful of biases, including the surprising, as proposed by State et al. (2013), migratory connections between the majority of countries and the United States. Secondly, while the delivery of data is extremely fast and relatively egalitarian in the sense of worldwide access, the data itself does not differ in its nature from other sources of information on migration flows. This is to say that a geolocated service user needs to arrive at the destination (e.g., to Spain) and log into the particular Internet service to be geotagged for a first time. Subsequently, the user must use it several more times during a period of time in order to be considered a short-time traveler or migrant.

Time Shifting of the Information Search

The main assumption of this study is that persons that plan any migratory activity would acquire knowledge about a potential destination through the media, and as a part of this process, they would use the Internet to search certain information related to the geographical destination that they aim to reach (Desiderio, 2012; Hamel, 2009). That assumption is similar to the ILI detection study (Ginsberg et al., 2009) described previously. Nevertheless, there would be some important modifications in the construction of the relation between real-life occurrence of events and a change in a query's popularity in Google Search. The occurrence of queries related to influenza was assumed to emerge as a result of an arising ILI epidemic (Ginsberg et al., 2009). In other words, people or relatives of people who detected symptoms would relatively immediately search for treatment advice and medications, prompting an increase in searches.

In the study of migration, the order is assumed to be the opposite. The change in the popularity of searches would precede the actual change in migration. Another crucial difference to the Ginsberg et al. (2009) study is a time lag between change in query popularity in Google and the change in migration flow occurrence. In fact, in the case of migration studies, this difference would be twofold. First, there is often a time lag between an immigrant's arrival at the destination and registering in the population register or being registered by some system of population control. Second, there should be a time lag between change in the popularity of search for migration related queries (e.g., job in the destination, flight tickets, apartment) and the change in migration flows itself. Since the studies on Internet use by the migrants focus mainly on practices in destination countries, there are no time estimations for a prospective immigrant to search online for information before the event of migration (Chen, 2010; De Tona & Whelan, 2009; Fox & Livingston, 2007; Kissau, 2009; Orozco, Burgess & Ascoli, 2010). For this study, we will assume a time shift of several months between the events when totaling the time lags. This assumption is concordant with migration models

that include pre-migration preparations (Benson-Rea & Rawlinson, 2003), in which early settlement is not until the fourth stage of a five-stage migration process.

The Spanish Residential Variations Statistic as an Exemplary Database

The assumption of time lag impacts data sources applicable for the analysis. The predicted gap of several months would make the use of annual migration flows data irrelevant, since these would not be able to precisely render the relationship between the change of searched query popularity and change of migratory flows. Unfortunately, most data on migration flows are published on an annual basis. One relevant option, however, is use of the anonymized micro-data of Residential Variation Statistics, based on Municipal Register (i.e., *padrón*) provided by the National Statistics Institute of Spain (esp. Instituto Nacional de Estadística [INE]). This dataset provides daily data on new registrations or address changes of Spanish residents in relation to countries of previous residence.

Furthermore, State et al. (2013) underscored the role of language and former colonial links in the migratory flows based on the use of IP geotagging. Following this pattern, this study analyzes the migratory flows between Latin America and Spain. The selection of Spain as a destination country is also supported by the growth of Spain's importance in receiving a large number of immigrants. Furthermore, the Americas constitute the origin of the majority of newly registered residents with previous residence outside Spain in the 21st century. It is especially worthwhile to point out that 34.19% of new registrations of residents with previous foreign addresses had arrived from the Americas, and that this number even slightly surpassed those with previous European residence (INE, 2012).

Still, it should also be stressed that Google Trends does not reliably cover the search traffic from all Latin American countries and, in certain cases, Google Search traffic volume may be too limited to be captured by the current database of Google Trends. Consequently, the results from only Spanish speaking countries that are fully covered by Google Trends during a particular time span are analyzed. The very similar limitation addresses particular queries that could be used in this research, but are not adequately covered for the whole time span of the study's interest, and therefore are eliminated from analysis.

Taking into account all the described assumptions and limitations, the hypothesis of the study is that the time-shifted change in proportion of migration-related queries made to Google Search in the sending country and reported by Google Trends predicts changes in the volume of international migration flows from sending country to destination.

Method

Identification and description of datasets

There are two sources of data used in this study: the time series of popularity of selected migration-related queries made to Google Search reported by Google Trends, and the Spanish Residential Variation Statistics based on Spanish Municipal Register (i.e., *padrón*). The data for both is examined in monthly intervals and covers a time span of six years from January 2005 to December 2010.

The google trends dataset

Google Trends is a publicly accessible tool that needs to be efficiently operated and does not provide raw levels of queries searched, but reports an index with data categorized at the national, regional and municipal level depending on the geographic region coverage requested. Also, the data accessible for Google Trends users is sampled from the entire Google Search database and limited to the queries with significant volume (Choi & Varian, 2009b). Choi and Varian (2011:3) offer precise description of the index construction:

> *The query index is based on query share: the total query volume for the search term in question within a particular geographic region divided by the total number of queries in that region during the time period being examined. The maximum query share in the time period specified is normalized to be 100 and the query share at the initial date being examined is normalized to be zero.*

The analyzed queries are intended to be related to the elements of international migration processes and their use depends on their coverage by the Google Trends database. In particular, the selection of queries studied here is based on previous literature on international migration (Anderson & Blinder, 2013; Castles & Miller, 1993; Colectivo Ioé, 2008; Demuth, 2000), and follows the logic presented in the previously introduced example of Choi and Varian (2009b). Therefore, at the preliminary stage of this study, several dozens of similar queries (e.g., *trabajo en España, trabajo España, empleo en España, visado España, visado para España, embajada de España, embajada España*) were retrieved from the Google Trends database for the countries of interest in order to select the queries with the most comprehensive geographic and longitudinal coverage.

This preliminary review revealed several limitations that were further applied to this study. First, the vast majority of the queries provided outlying results for the year 2004. This can be explained with the year 2004 as being the first year for which Google Trends collected data, and therefore some lower volume queries might have been related to shortcomings in coverage. Furthermore, by January 1, 2011, there was a significant change introduced in the standards of data exploitation by Google Trends, which may result in discontinuity with previous longitudinal data.[i] Based on these assumptions, the actual analysis has been conducted on data that covers a time span of six years, from January 2005 to December 2010.

Additionally, the above mentioned preliminary exploration of queries regarded the geographical coverage for particular Latin American countries. The list of countries that has been checked against Google Trends data availability has been constructed from the summed *padrón* statistics of the new registrations of Spanish inhabitants that previously resided in Spanish speaking Latin American countries (INE, 2012). The data covered the period of research interest between the years 2005 and 2010. The results of this preliminary analysis demonstrated that, as mentioned by other authors (e.g. Choi & Varian 2009a), Google Trends still carries out significant shortcomings, especially in relation to data provided for developing (here, Latin American) countries. In order to overcome these Google Trends drawbacks, only three queries based on the criteria mentioned in the previous paragraphs — and that are simultaneously covered by Google Trends for three, large migration sending countries — have been chosen for further analysis.

In particular, this study analyzes the queries for Colombia, Argentina and Peru that respectively were positioned as second, third and fourth in the list of Spanish-speaking Latin American countries that constituted previous residence for newly registered Spanish inhabitants during the period of interest. More specifically, these countries are reported as

[i] https://support.google.com/trends/answer/1383240?hl=en

previous residences for almost 30% of all newly registered inhabitants with previous Latin American addresses, and almost 9% of all newly registered inhabitants with previous foreign addresses for this period (INE, 2012).

In conclusion, the selection is related to the particular elements of migration processes and it additionally depends on the shortcomings of the Google Trends database's coverage. The particular queries selected for analysis in this study are: *trabajo en España* (work in Spain) as related to economic migration process, *embajada de España* (Spanish Embassy) as related to Spanish (and EU) visas and entry requirement and simply *España* (Spain) based on previous findings in the similarity between tourism and migration related global network activity (State et al., 2013).

Still, one could argue that the more appropriate queries would be ones related to search for employment in a particular municipality. This is a legitimate point taking into account the geographic precision in the practice of employment seeking. In this case, particular examples could be related to the Spanish gateway cities for immigrants (Colectivo Ioé, 2008), e.g., *trabajo en Barcelona* or *trabajo en Madrid*. Unfortunately, Google Trends does not provide longitudinal data for these queries that satisfy the time range in order to carry out the comparison. This however, does not invalidate the use of the query *trabajo en España*. There are two premises significant for this assumption. First, based on logic, the more general query might be used at the beginning of the job-seeking process in order to identify general resources (e.g., internet job boards or classified ads) provided at the national level. In this case, the process of geographic narrowing would be next in line. This logic is supported by the statistical premise provided by Google Trends itself. The mere fact that coverage is provided for the more general query vs. more locally narrowed one indicates its higher popularity in Google Search, and therefore relevance for this study.

Spanish municipal register (i.e. padrón)

According to Spanish Law[ii] "… *pardón* is an administrative register which consist of the neighbors of the municipality…" Each and every individual that resides in a municipality is required to enroll under the appropriate penalties if this obligation is not fulfilled. The first and basic function of the *pardón* is to recount the number of inhabitants of each Spanish municipality (García-Pérez, 2007). For this study in particular, it is crucial that every person that enrolls in the *padrón* is required to provide a location of his or her previous municipality or country of residence (Suero-Salamanca, 1999). The latter allows for the tracking of international migration flows on the basis of the *padrón* data.

The second function of the *padrón* is for the individual to be able to demonstrate the residence in the municipality if needed. This function is met through the certifications elaborated from the *padrón* database. These are public documents with administrative power (García-Pérez, 2007). From the perspective of the individual, being enrolled within a *padrón* also provides some particular privileges like the right to use public services present in the municipality (Suero-Salamanca, 1999).

It is important to restate some of the characteristics of a *padrón* that make it a valid source of information on migration flows. In contrast to some registration systems, the *padrón* encourages immigrants to undergo an enrollment procedure. First of all, it is accessible for all the inhabitants with no prejudice or status check for irregular immigrants. Second, before the year 2012, the enrollment gave immediate access to free public health care regardless of

[ii] el art. 16 de la Ley 7/1985 de 02 de abril, reguladora de las bases del Régimen Local, modificada por la Ley 4/1996 de 10 de Enero11

immigration status. More recently, this has been locally limited due to the economic recession. Third, the enrollment might be used as further documentation of seniority of residence in order to undergo processes related to residence, nationalization or family rejoining. What an immigrant needs in order to register is a valid document (e.g., passport) and a relatively easily accessible apartment lease.

This research makes use of the anonymized micro-data of Residential Variation Statistics based on *padrón*, and provided by INE. This data set provides information about each particular event of an individual's registration in each and every community with more than 10,000 inhabitants in Spain. Data for the towns and villages with less than 10,000 people are additionally anonymized by not providing the community name but only the region name in order to ensure data anonymity.

The original daily data of individual enrollments in the Residential Variation Statistics has been aggregated into the monthly intervals in order to facilitate comparison with the data retrieved from the Google Trends web-based application. The original data set of 15,684,892 data points that represent each enrollment event that occurred in the given time span in Spain (2005: 2,357,656; 2006: 2,715,449; 2007: 2,980,684; 2008: 2,635,679; 2009: 2,475,632; 2010: 2,519,792) has been narrowed down to the data points representing individuals with direct previous residence in Colombia, Argentina or Peru. These data were subsequently aggregated into the monthly data for each country. The aggregate data includes 72 monthly intervals (aggregated data points) for each analyzed country.

Statistical analysis of time series data

Following the indications of the previous research and literature, this study applies the Cross-Correlation Function on the pre-whitened time series data in order to assess the relation between Google Search and Residential Variation Statistics at various monthly lags of the predictor. Furthermore, regression models are examined in order to assess the predictability at the particular lags that result to be significantly correlated (Choi & Varian, 2009a; National Cooperative Highway Research Program Report, 1997; McCleary & Hay, 1980; Tabachnick & Fidell, 2010).

Specifically, the change of the proportion in each of the Google Search queries has been compared to the change in proportion of the number of residents enrolled in the *padrón* for which the previous residence was the particular country of origin from the Google Trends data. This resulted in nine comparison pairs. Four stages of the analysis have been conducted on each pair of the time series: (a) The visual analysis of the Cross-Correlation Function (CCF) and sequence charts of the original non-stationary data that show the correlation between the general trends of the time series while taking into account possible lags in reaction of *padrón* to the Google Search queries changes. Their statistical significance is not formally assessed due to the possible lack of data randomness; (b) The analysis of time series' stationarity based on the visual inspection of Auto-Correlation Function (ACF) and Partial Auto Correlation Function (PACF) has been performed. If the series was found to be non-stationary, it was stationarized with the use of first order differentiation (that occurred enough in order to stationarize these particular time series according to the repeated PACF and ACF results).

Therefore, instead of the direct comparison of equivalent months, a comparison of the difference in the number of immigrants enrolled in *padrón* to Google Search queries for the particular month (t) and the previous month (t-1) have been assessed. The transformed time series has been used in further analysis; (c) The CCF has been again performed on the stationarized time series in order to evaluate actual correlation between two time series and find the most highly correlating months' lag; (d) The lags that were shown to be significantly

correlated by the CCF and that were in accordance with the previous theoretical assumption have been assessed for the predictability with the Linear Regression model. The Durbin-Watson statistics has been additionally assessed to examine if autocorrelation of the time series has not influenced the regression results.

Results

Colombia. In the case of Colombia, the visual inspection of the time sequence chart showed a strong downward trend in all the predictors analyzed. The DV's trend was more diverse, but would follow the same downward trend if the lag was applied to the predictors. This again indicates strong non-stationarity of the time-series in question. CCF performed on the non-stationary time series showed that *trabajo en España*, *embajada de España*, and *España* show the strongest correlation with the DV at a lag of nine months applied to the predictors. Since the ACF and PACF confirmed the series to be non-stationary, no conclusion could be inferred on the significance of the correlation coefficient for these time series before the first order differentiation that was further applied and evaluated with repeated ACF and PACF. Afterwards, the CCF performed on the stationarized time series again indicated the significant correlation with the Residential Variation data on lag nine for the *trabajo en España*, *embajada de España*, and *España*.

Furthermore, all three queries that were tested for their nine months lags have predictive power. The *trabajo en España* [R= .292, R^2 = .085, adj.R^2 = .070, F(1, 60) = 5.597, p = .021] and *embajada de España* [R= .348, R^2 = .121, adj.R^2 = .107, F(1, 60) = 8.276, p = .006] occurred to significantly explain small proportion of variance in the Residential Variation, while the *España* [R= .459, R^2 = .211, adj.R^2 = .198, F(1, 60) = 16.037, p = .000] occurred to significantly explain small to moderate proportion of variance in the Residential Variation (see Table 1).

Table 1: Regression models predicting Spanish Residential Variation of immigrants with previous residence in Colombia from *trabajo en España*, *embajda de España* and *España* Google searches performed in Colombia

	Residential Variation	
Variable	B	β
Constant	-3.797	
Trabajo en España	9.693*	.292
R^2	.085	
F	5.597*	
Constant	16.650	
Embajada de España	18.415**	.348
R^2	.121	
F	8.276**	
Constant	26.029	
España	33.467***	.459
R^2	.211	
F	16.037***	

Note. *p <.05, **p<.01, *** p< .001

Argentina. In the case of Argentina, the visual inspection of the time sequence chart showed a strong downward trend in all the time series analyzed. This indicates a similarity of the general trend over time between the predictors and the outcome variable, but also the strong non-stationarity of the time-series. CCF performed on the non-stationarized time series confirm that the series are highly correlated. The queries *trabajo en España* and *embajada de*

España showed the strongest correlation with the Residential Variation at lag 2 of both predictors. The query *España* showed the strongest correlation with the Residential Variation at lag 0 of both predictors. Since the ACF and PACF confirmed the series to be non-stationary, no conclusion could be inferred on the significance of the correlation coefficient for these time series before the first order differentiation that was further applied and evaluated with repeated ACF and PACF. Afterwards, the CCF performed on the stationarized time series indicated a significant correlation with the Residential Variation data on lag 4 for the *trabajo en España* and on lag 8 for the *embajada de España*. The query *España* did not show significant correlation with the Residential Variation data.

In order to statistically assess the predictive power of the queries made to Google Search on the international migration flows, linear regression was performed on the lags of the stationarized IVs that showed to be significantly correlated according to the CCF results. The query *trabajo en España* lagged by four months [R= .360, R2 = .130, adj.R^2 = .116, F(1, 65) = 9.696, p = .003], and the query *embajada de España* lagged by eight months [R= .389, R^2 = .151, adj.R^2 = .137, F(1, 61) = 10.869, p = .002] appeared to significantly explain a small to moderate proportion of variance in the Residential Variation (see Table 2).

Table 2: Regression models predicting Spanish Residential Variation of immigrants with previous residence in Argentina from trabajo en España and embajada de España Google searches performed in Argentina

	Residential Variation	
Variable	B	β
Constant	-8.086	
Trabajo en España	9.980**	.360
R^2	.130	
F	9.696**	
Constant	-8.609	
Embajada de España	11.459**	.389
R^2	.151	
F	10.869**	

Note. *p <.05, **p<.01, *** p< .001

Peru. In the case of Peru, the visual inspection of the time sequence chart again showed a strong downward trend in all the predictors analyzed and more diverse trend of DV that would follow the predictors' downward trend if the lag was applied to the predictors. This again indicates strong non-stationarity of the time-series in question. The CCF performed on the non-stationary predictors showed the strongest correlation with the DV at lag eight for *trabajo en España*, lag nine for *embajada de España* and lag 12 for *España*. Since the ACF and PACF confirmed the series to be non-stationary, no conclusion could be inferred on the significance of the correlation coefficient for these time series before the first order differentiation that was further applied and evaluated with repeated ACF and PACF. Afterwards, the CCF performed on the stationarized time series indicated no significant correlation for queries *trabajo en España* and *España*.

A significant correlation with the Residential Variation data on lag nine for the *embajada de España* was found. Therefore, the only tested query was *embajada de España* lagged by nine months in respect to the Residential Variation. The regression results showed this query to significantly explain small proportion of variance in the Residential Variation [R= .281, R^2 = .079, adj.R^2 = .064, F(1, 60) = 5.147, p = .027] (see Table 3).

Table 3: Regression models predicting Spanish Residential Variation of immigrants with previous residence in Peru from *trabajo en España* Google searches performed in Peru

	Residential Variation	
Variable	B	β
Constant	9.626	
Trabajo en España	16.274*	.281
R^2	.079	
F	5.147*	

Note. *p <.05, **p<.01, *** p< .001

Conclusions and Discussion

The results of the regression analysis partially support the hypothesis that the time-shifted change in proportion of migration-related queries made to Google Search in sending country and reported by Google Trends predicts changes in volume of international migration flows from sending country to destination. The differences of the results between countries show that a number of particular migration flow characteristics should be taken into account while building the Residential Variation prediction models based on the queries to Google Search reported by Google Trends.

While the data for Colombia and Peru showed the correlation of stationarized time series at lags nine and eight months respectively, the data for Argentina showed more diverse results. Further exploratory analysis of the Residential Variation datasets revealed an important difference between the structure of the inhabitants that arrived from Argentina, and those that arrived from Peru and Colombia. The high proportion (up to 40 – 50%) of those that enrolled after arriving from Argentina were of European nationality. That fact might importantly influence not only the travel and employment requirements applied to them, but also the requirements applied to the families that would further join them. Concomitantly, the time patterns of lags between search for particular queries (e.g. search for *embajada de España*) would be irrelevant to these immigrants. This condition in the case of Argentinean immigrants has already been found in previous research conducted in Spain (Morén, Mas & Wladyka, 2016). Further studies should consider controlling for this and similar factors.

The relatively weak or at best moderate predictability of changes in the Residential Variation could be related to the average rate of internet penetration during the researched period. Because of the scarcity and inconsistency of data on Internet penetration from developing countries, it is difficult to statistically test this assumption. Still, some preliminary indications may be provided here as a starting point for the future research. The predictability of results given for Peru is shown to be less powerful than that for Colombia and Argentina. Accordingly, the International Telecommunication Union (2016) states that at the end of the researched period in the year 2010, the percentage of individuals using the Internet was lowest in Peru (34.77%), as compared to Colombia (36.50%) and Argentina (45%). On the other hand, the average percentage of individuals using the Internet for the entire researched period between 2005 and 2010 provides slightly different picture: Colombia (23.37%), Argentina (28.61%) and Peru (26.62%) (International Telecommunication Union, 2016). It is worth noting that while there is no correlation between the predictability and the Internet penetration rate between analyzed countries, the Internet penetration rate is an interesting variable to include in further studies, as the applicability of Google Search queries for predicting the Residential Variation depends highly on Internet expansion in developing countries.

On the other hand, this study shows the fragility of Google Search predictability interpretation for extraneous factors (e.g. citizenship). Also, other variables like language barriers and migration type should be considered in further studies. In particular, the discussed predictability of migration flows could be limited (or at least in need of development of alternative approaches) in the case of urgent movements observed by refugees, for example, fleeing from events like natural disasters or conflicts. Moreover, the nature of the tool itself (looking at changes of query popularity) would be cumbersome in tracking less numerous migratory patterns, like retirement migrants and affluent sun seekers, especially crossing borders of regions with varying languages.

As previously explained, there has been no research done on the pre-migratory web-searching habits by individuals that undertake migration processes. Therefore, an interesting step in future studies could be an introduction of qualitative research methods, based on the semi-structured interviews, that could shed light on what kind of information and in what ways immigrants search online about their destination country. This qualitative data could be used to generate potential queries to Google Search in order to find information about the destination country and city. Furthermore, inclusion of additional, more diverse queries as well as languages, could possibly provide a basic differentiation between migrant typologies and should be further evaluated.

Although many challenges exist in the construction of an international migration flows prediction model based on Google Search queries, this study supports the assumption that data on changes in Google Search query popularity obtained from Google Trends might constitute a useful and at least additional tool in predicting the changes in migration flows several months in advance. Furthermore, the development of Google Trends (or a similar tool), especially in the sense of coverage for developing countries and less popular queries (e.g., focused on particular municipalities), and the growth of Internet penetration rate therein, are the factors that could convert Google Trends into a multipurpose tool in estimating and predicting international migration flows.

From the policymakers' standpoint, to address issues and challenges brought about by migration, there is often a need to provide long-term solutions that involve collaboration among local and national levels of administration. Thus, the use of this tool could help local authorities adjust their policies like housing and welfare, as well as adjust the expected demand for services in order to meet the social reality produced by an immediate future of immigration flows. It could also raise awareness about the current demographic situation of the region or city in question. In particular, an insight into geographic origins of potential immigrants would allow authorities to train appropriately skilled translators and social cultural mediators.

Moreover, the data could also be of use for sending countries, since the migrants' remittances directly influence the economy of origin. The impact of remittances can be crucial for sustaining the robustness of the country's economy, as, for example, during the global economic crisis (UN, 1998). The recent economic recession, during which a European Union member — Poland — did not formally enter the economic recession (lack of Gross Domestic Product decrease), is a perfect example of such an impact. The hundreds of thousands of Polish citizens that migrated to more developed European countries (mainly the UK, Ireland and Germany) not only strengthened the Polish economy with remittances, but also alleviated issues related to unemployment on the Polish labor market noted during the European recession (Ministerstwo Spraw Zagranicznych, 2013). While this positive effect contrasts with the negative "brain drain" phenomenon (Krajowy Punkt Kontaktowy Europejskiej Sieci Migracyjnej, 2011), both demonstrate the economic significance of international migration flows for countries of origin.

Conversely, in tackling the political implications for the countries of origin, one could raise concern regarding the ethical implications of gauging the spatial mobility of migrants, especially those fleeing from persecution or war. The use of Google Trends evaluated in this article assumes processing of only anonymized data that should not bring privacy concerns and is limited in predictability of urgent population movements. However, any additional insight into the movements of populations fleeing from armed conflicts could potentially be of a malevolent political use for a country of origin or tactical use for parties engaged in the particular conflict, so should be carefully evaluated and further discussed.

Acknowledgments

The author would like to thank Dr. Bernardo De La Garza, Dr. Matthew Johnson and Dr. William Yaworsky for their valuable suggestions in the preparation of this study. This article is linked to the MA thesis prepared by the author at the University of Texas at Brownsville and directed by Dr. William Yaworsky.

References

Anderson, B. & Blinder, S. (2013). *Who counts as a migrant? Definitions and their consequences*. London: The Migration Observatory at the University of Oxford. Retrieved from http://www.migrationobservatory.ox.ac.uk/sites/files/migobs/Briefing%20-%20Who%20Counts%20as%20a%20Migrant_0.pdf

Benson-Rea, M. & Rawlinson, S. (2003). Highly skilled and business migrants: Information processes and settlement outcomes. *International Migration,* 41(2), 59-79. DOI: http://dx.doi.org/10.1111/1468-2435.00235

Castles S. & Miller, M. J. (1993). *The age of migration: International population movements in the modern world*. New York: The Guilford Press.

Chen, W. (2010). Internet-usage patterns of immigrants in the process of intercultural adaptation. *Cyberpsychology, Behavior, and Social Networking,* 13(4), 387-399. DOI: https://doi.org/10.1089/cyber.2009.0249

Choi, H. & Varian, H. (2009a). *Predicting initial claims for unemployment benefits*. Technical report. Retrieved from http://research.google.com/archive/papers/initialclaimsUS.pdf

Choi, H. & Varian, H. (2009b). *Predicting the present with Google Trends*. Retrieved from http://google.com/googleblogs/pdfs/google_predicting_the_present.pdf

Choi, H. & Varian, H. (2011). *Predicting the present with Google Trends*. Retrieved from http://people.ischool.berkeley.edu/~hal/Papers/2011/ptp.pdf

Colectivo Ioé. (2008). *Inmigrantes, nuevos ciudadanos: ¿Hacia una España plural e intercultural?* Madrid: Fundación de las Cajas de Ahorros (FUNCAS).

ComScore. (2011). *Google sites accounts for 9 of 10 searches conducted in Latin America*. Retrieved from http://www.comscore.com/Insights/Press_Releases/2011/5/Google_Sites_Accounts_for_9_of_10_Searches_Conducted_in_Latin_America

De Tona, C. & Whelan, A. (2009). 'Re-mediating' the ruptures of migration: The use of internet and mobile phones in migrant women's organisations in Ireland. *Translocations: Migration and Social Change.* Retrieved from http://ec.europa.eu/ewsi/UDRW/images/items/docl_10859_479879574.pdf

Demuth, A. (2000). Some conceptual thoughts on migration research. In B. Agozino (Ed.), *Theoretical and methodological issues in migration research* (pp. 22-26). Aldershot, England: Ashgate Publishing Ltd.

Desiderio, M. V. (2012). *Practices of access to labour market information for migrants and employers: An overview LINET 2012 research findings*. Independent Network of Labour Migration and Integration Experts. Retrieved from www.labourmigration.eu

FairSearch (2013). *FairSearch Fact Sheet*. Retrieved from http://www.fairsearch.org/wp-content/uploads/2011/06/Draft-Core-FairSearch-Fact-Sheet-051812.pdf

Ferrari, L., Rosi, A., Mamei, M. & Zambonelli, F. (2011). *Extracting urban patterns from location-based social networks*. Paper presented at the 3rd ACM SIGSPATIAL International Workshop on Location-Based Social Networks, Chicago, IL. DOI: http://dx.doi.org/10.1145/2063212.2063226

Fox, S. & Livingston, G. (2007). *Latinos online: Hispanics with lower levels of education and English proficiency remain largely disconnected from the internet*. Washington, DC: Pew Hispanic Center and Pew Internet Project. Retrieved from http://www.pewhispanic.org/files/reports/73.pdf

García-Pérez, Mª. S. (2007). El padrón municipal de habitantes: Origen, evolución y significado [The municipal register of inhabitants: Origin, evolution and meaning] Hispania Nova. *Revista de Historia Contemporánea* [Journal of Contemporary History]. 7. Retrieved from http://hispanianova.rediris.es

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), 1012-1014. DOI: https://doi.org/10.1038/nature07634

Google Trends. (2012). *Analyzing Data*. [Data file]. Available via: http://support.google.com/trends/topic/13975?hl=en&ref_topic=13762

Gueye, B., Uhlig, S. & Fdida, S. (2007). *Investigating the imprecision of IP block-based geolocation*. Paper presented at PAM 2007, Louvain-la-Neuve, Belgium. Retrieved from http://www.eecs.qmul.ac.uk/~steve/papers/Geolocation-pam07.pdf

Hamel, J. Y. (2009). Information and communication technologies and migration. *Human Development Research Paper 39*. New York: United Nations Development Programme, Human Development Report Office. Retrieved from http://hdr.undp.org/en/reports/global/hdr2009/papers/HDRP_2009_39.pdf

Hui, P., Mortier, R., Piorkowski, M., Henderson, T. & Crowcoft, J. (2010). *Planet-scale human mobility measurement*. Paper presented at the second ACM International Workshop on Hot Topics in Planet-Scale Measurement (HotPlanet). New York, NY. DOI: http://dx.doi.org/10.1145/1834616.1834618

Instituto Nacional de Estadística. (2012). Data extracted from the census database for 1st of January of 2012. Retrieved from http://www.ine.es

International Telecommunication Union. (2016). *Percentage of individuals using the Internet*. Retrieved from http://www.itu.int/en/ITU-D/Statistics/Pages/stat/default.aspx

King, K. F. (2010). Geolocation and federalism on the Internet: Cutting Internet gambling's Gordian Knot. *The Columbia Science and Technology Law Review*, 11(41). Retrieved from http://www.stlr.org/cite.cgi?volume=11&article=2

Kissau, K. (2009). Online spheres of migrants in Germany: How and why migrants use the Internet. In C. Dietz and P. Stammen (Eds.), *Media on the move: Migrants and minorities and the media* (pp. 80-84). 4th Symposioum on Forum Media and Development, Bonn 2008, Aachen Germany, Catholic Media Council (CAMECO). Retrieved from http://www.cameco.org/files/mediaonthemove-kissau.pdf

Krajowy Punkt Kontaktowy Europejskiej Sieci Migracyjnej (2011) *Migracja tymczasowa i cyrkulacyjna w Polsce: dotychczasowe doświadczenia, uregulowania prawne i opcje na przyszłość Lata 2004-2009* [Temporary and circular migration in Poland: Past experiences, legal regulations and future options. Years 2004-2009]. Retrieved from http://ec.europa.eu/dgs/home-affairs/what-we-do/networks/european_migration_network/reports/docs/emn-studies/circular-migration/pl_study_on_temporary_and_circular_migration_pl_version_pl.pdf

Kupiszewska, D. & Nowok, B. (2008). Comparability of statistics on international migration flows in the European Union. In J. Raymer & F. Willekens (Eds.) *International migration in Europe: Data, models and estimates* (pp. 41-71). Chichester, England: John Wiley & Sons.

McCleary, R. & Hay, R. (1980). *Applied time series analysis for the social sciences.* London: Sage Publications.

Ministerstwo Spraw Zagranicznych (2013) *Społeczno-gospodarcze efekty członkostwa Polski w Unii Europejskiej (1 maja 2004 – 1 maja 2013).* Główne wnioski w związku z dziewiątą rocznicą przystąpienia Polski do UE [Socioeconomic effects of Poland's membership in the European Union (May 1, 2004 – May 1, 2013): Main conclusions on the ninth anniversary of Poland's assession to the EU]. Retrieved from http://www.msz.gov.pl/resource/94e4e616-9554-4c76-af2e-6615e3196f83:JCR

Mohebbi, M., Vanderkam, D., Kodysh, J., Schonberger, R., Choi, H. & Kumar, S. (2011). *Google correlate whitepaper.* Retrieved from https://www.google.com/trends/correlate/whitepaper.pdf

Morén-Alegret, R., Mas, A. & Wladyka, D. (2016) Inter-group perceptions and representations in Barcelona. A comparison of Poble Sec and Sagrada Familia neighbourhoods. In I. Ponzo & F. Pastore (Eds.) *Inter-group relations and migrant integration in European cities. IMISCOE Research Series* (pp. 89-121). Cham: Springer.

Moses, J. W. (2012). EMIG 1.2: A global time series of annual emigration flows. *International Migration*, 53(5), 47-60. DOI: http://dx.doi.org/10.1111/imig.12026

National Cooperative Highway Research Program Report. (1997). *A guidebook for forecasting freight transportation demand* (Report No. 388). Washington, D.C: Transportation Research Board.

Orozco, M., Burgess, E. & Ascoli, N. (2010). Is there a match among migrants, remittances and technology*? Inter-American Dialogue,* September 2010. Retrieved from http://www.thedialogue.org/PublicationFiles/a%20match%20in%20migrants%20remittances%20and%20technology%20MO_FINAL_11.4.101.pdf

State, B., Weber, I. & Zagheni, E. (2013). *Using IP address geolocation to analyze international migration and mobility patterns*. Paper presented at the sixth ACM international conference on Web search and data mining, Rome, Italy. DOI: http://dx.doi.org/10.1145/2433396.2433432

Suero-Salamanca, J. A. (1999) Estudio sobre el padrón municipal de habitantes [Study on the municipal register of inhabitants]. *Actualidad Administrativa* [Administrative News], 15(12). Retrieved from http://hispanianova.rediris.es/7/articulos/7a005.pdf

Svantesson, D. J. B. (2004). Geo-location technologies and other means of placing borders on the 'borderless' Internet. *Journal of Computer & Information Law*, 23(1), 101-139. Retrieved from http://epublications.bond.edu.au/law_pubs/63

Tabachnick, B. & Fidell, L. (2010). *Using multivariate statistics* (5th ed.). Boston: Pearson Education.

United Nations. (1998), *Recommendations on statistics of international migration*: *Revision 1*. New York: United Nations. Retrieved from https://unstats.un.org/unsd/publication/SeriesM/SeriesM_58rev1e.pdf

Zagheni, E. & Weber, I. (2012). *You are where you e-mail: Using e-mail data to estimate international migration rates*. Paper presented at the 3rd Annual ACM Web Science Conference, New York, NY. DOI: http://dx.doi.org/10.1145/2380718.2380764