Desaparecidxs: leveraging data on disappeared persons using Twitter

Extended abstract submitted to the European Population Conference 2020

Diego Alburez-Gutierrez
Max Planck Institute for Demographic Research
Rostock, Gemany

Carolina Coimbra Vieira
Max Planck Institute for Demographic Research
Rostock, Gemany

Marília R. Nepomuceno
Max Planck Institute for Demographic Research
Rostock, Gemany

Tom Theile
Max Planck Institute for Demographic Research
Rostock, Gemany

ABSTRACT

Demographers have long neglected the population of disappeared. First, because it can represent a small proportion of the total population, especially in high-income countries. Second, because the data of disappeared persons are not always easily accessible. However, the representativeness of the number of disappeared persons can be important in regions that are experiencing dictatorships or are in violent settings. To date, little is known about the composition of this population by sex, age, or race/ethnicity. This study aims to use social media data to investigate the population of disappeared in Latin America, a region where several types of violence are increasing. By combining different data sources, including the Twitter data, our analysis offer a detailed picture of the population of disappeared in Guatemala. Finally, given the quality of the data shared on Twitter, this work can be replicated to other countries (e.g.Brazil) where data at individual level is available.

1 INTRODUCTION

Disappeared persons inhabit a liminal state - it is unknown whether they are dead or alive. Disappearances are a common phenomenon in dictatorships and violent conflict zones but they also take place outside these settings. People can disappear because of violent causes (homicides, domestic violence, sexual violence, child abuse, elder abuse) and non-violent ones (mental disorders, flight, drug abuse). The reasons why an individual disappears may also be associated with age, gender, race, socioeconomic level, and social norms. However, little is still known about the characteristics of the disappeared persons. We are interested in disappearances in the context of Latin America, where violence has increased recently, especially among population living in vulnerable areas (McIlwaine and Moser 2001; Soares-Filho 2011; Meneghel and Hirakata 2011; Garcia and Aburto 2019).

Demographers have neglected the population of disappeared persons. First, because it can represent a small proportion of the total population, especially in high-income countries. Second, because the data of disappeared persons are not always easily accessible. However, there are important reasons that justify focusing on this population. On the one hand, disappeared individuals are not registered as a death in vital statistics systems. This may affect the estimations of sub-national demographic rates, particularly if the increasing violence disproportionately affects certain population subgroups, such as women, children, or some specific race/ethnicity. On the other hand, the phenomenon of disappearances may have profound implications for individuals and societies as a whole. The disappearance of a family member is a highly stressful event. Previous studies have shown that the death of a relative can increase the risk of death, and can also result in severe mental health consequences of those left behind (Hendrickson 2009; Li et al. 2003; Rostila et al. 2012). The phenomenon of disappearance could be even more distressing for families than the death *per se* due to the uncertainty in whether a person disappeared is dead or alive, but no studies on the subject in the demographic literature.

Our study aims to conduct a comparative analysis of the population of disappeared persons using data from Guatemala and (at a later stage) Brazil. Studying this population can be further complicated by the difficulty of accessing statistics on disappeared people broken down by age and sex. The relevant authorities might refuse this information on different grounds including privacy concerns, protection of youth or unavailability of data. In this paper, we explore an alternative data source for obtaining real-time demographic data on disappeared persons from a popular social network, Twitter. We show how Twitter data can complement official statistics produced by police departments, especially in cases where only aggregate data is available. This is the case in Guatemala, where the National Police only provides information of disappeared persons by four age groups (0-17, 18-35, 36-65 and 56+). We overcome this limitation by collecting data from the official Twitter account of the Alerta Alba-Keneth (*@alba_keneth*), a governmental warning system tasked with disseminating information about disappeared youth (individuals aged 0-17) in Guatemala.

In this paper, we introduce a methodology for collecting the relevant data from Twitter in a systematic way, including sophisticated image processing techniques developed for this purpose. From the media shared by this account on Twitter we determined the age, sex, and skin color of more than 2,000 disappeared youth, and date and location of the disappearance for the period October 2018-August 2019. By combining two data sources we obtained a detailed picture of the demographic dynamics of disappeared persons. In this abstract we show results from Guatemala, however we aim to expand the analysis for Brazil, where information of disappeared persons are available at the individual level for some states. This analysis will allow us to compare Twitter with official government data in order to show the efficiency and accuracy of data shared on Twitter.

2 DATA AND ANALYSIS

2.1 Data from the National Guatemalan Police

Disappeared youth in Guatemala (individuals under 18) can be reported to three government offices: the National Civilian Police (Policía Nacional Civil), the Prosecutor General's Office (Ministerio Público), or the Attorney General's Office (Procuraduría General de la Nación). Each of these offices produce their own statistics, which means that reports of disappearances are often duplicated across sources. In our experience, accessing these data is very difficult given the unwillingness of the relevant authorities to share it.

We obtained data on disappeared persons in Guatemala via multiple Freedom of Information Requests to the National Civilian Police. The data includes the number of disappearances reported to the police by age group and month of the event. The police department did not share individual-level data on the disappearances or information on the composition of the population by single-year age, gender, or ethnicity, as well as any details about the circumstances of the events. The data covers the period from January 2003 to August 2019, but we only used data after 2010 given data quality concerns for the earlier records. It is worth noting that these data do not cover all the reported cases of disappearances, as explained above, but only those reported to the police. In particular, we expect the police data to under-represent the total number of disappeared underage individuals in the country. We are currently preparing further Freedom of Information Requests to obtain data from other sources.

2.2 Twitter data

Twitter is a microblogging platform that allows users to express themselves and to record their thoughts in 280 characters at maximum. We can also describe Twitter as a social network since it is possible to follow users to be updated about their tweets. Tweets is the name of this short messages that a user can write and it may also contain photos, GIFs, videos, and links.

Due to the fact that Twitter is one of the most popular social media platforms with more than 300 million active users per month[1] people use this social network to tweet and share information with a wide audience. News outlets, academics, and governments use Twitter to share official information. We are particularly interested in the latter's use of Twitter to share information on disappeared people. The data provided by Twitter has been used extensively by researchers in social science (McCormick 2017). Recently, Tsoi et al. (2018) evaluates the effectiveness of using Twitter to search for people who got lost due to dementia.

We are interested in exploring the use of Twitter by the Guatemalan government to gather data on disappeared youth. We focus on the Alerta Alba-Keneth, the account of an inter-governmental agency tasked with the search, location and immediate protection of disappeared or abducted children and adolescents. The Alerta Alba-Keneth was established in 2010 to coordinate the efforts of the National Civilian Police, the Prosecutor General's Office, and the Attorney General's Office in locating disappeared youth. As such, the alert system benefits from data collected by the three government offices. The Alerta Alba-Keneth Twitter account, set up in 2015, has high visibility with almost 10 thousand followers. The *@alba_keneth* profile tweets images with information about new disappeared kids on a daily basis, following the same structure of the media shown in Figure 1.

The tweets were collected through an Application Programming Interface (API). Due to the configuration of the Twitter API[2], only the 3,200 most recent tweets from Alerta Alba-Keneth[3] could be collected. Of those, 3,054 contained media referring to a disappeared child. This number refers to unique individuals. All duplicated individuals have been removed, considering the fact that some images are tweeted more than one time. Finally, all the media collected have been pre-processed in order to extract information related to demographic attributes of the sample.[4] Privacy was a major concern when collecting the Twitter data. In spite of the fact that the data is publicly available, our study only discussed aggregate measures and includes no personally-identifying information of any disappeared person.

2.3 Skin color detection

We are interested in extracting the skin color from the images of the disappeared in order to determine the composition of the population of disappeared people by skin color - a proxy for ethnicity in Latin America (Telles et al 2014). Most of the images shared by the Alerta Alba-Keneth on Twitter include (self-reported) skin color but this information is not commonly available in other sources. In order to determine the skin color of the disappeared persons reported by the Alerta Alba-Keneth Twitter profile, we initially extracted the photographs of the youth from the media shared in the tweets (Figure 1). Then we detected the position of the face using Haar feature-based cascade classifiers (Sharifara et al 2014) and used k-means clustering on the RGB-color values of the cropped-out face with 3 clusters (Figure 2). We took the mean color of the largest cluster as the photographed skin color (the estimated skin color of the person photographed). We obtained standardized skin color by comparing the photographed skin color to the *Perla* palette, a classification of 11 skin-colors developed for social science research in Latin America (Telles et al 2014). The photographed skin-color was matched to the visually nearest Perla color using a measure of euclidean distance.

3 PRELIMINARY RESULTS: DISAPPEARED YOUTH IN GUATEMALA

In this section we focus on minors or youth, defined in Guatemala as the population under 18 years of age. According to the National Police, this age group represented more than half of the known cases of disappeared people in the country between 2011 and 2019 as shown in Figure 3.

We initially compare the data on disappeared youth across our two sources: the National Police and Alerta Alba keneth Twitter data. Figure 4 shows them side-by-side for the period in which both

[1] https://s22.q4cdn.com/826641620/files/doc_financials/2019/q1/Q1-2019-Slide-Presentation.pdf

[2] https://developer.twitter.com/en/docs/tweets/timelines/api-reference/get-statuses-user_timeline

[3] Official website available here: https://www.albakeneth.gob.gt/

[4] First, the crop() function of the image class in the Python Image Processing library, Pillow, is applied to crop the images. After this, we make use of Pytesseract, a Python Library for optical character recognition, to extract the text from images and information like age, skin color, and place and date of disappearance.

Figure 1: Example of the media in an Alerta Alba-Keneth shared on Twitter.

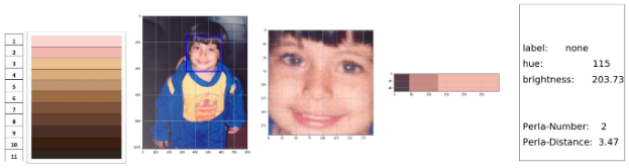

Figure 2: Graphic summary of the skin-color detection algorithm used on images of disappeared youth shared by the Alerta Alba-Keneth on Twitter (one of the authors is portrayed, as a child).

data sources overlap (October 2018 to August 2019). The number of disappearances is much higher from Twitter data when compared to the National Police data. The main reason is that the Twitter data contains information from two data sources: the National Police, and two other government agencies tracking disappeared youth. Despite the differences in the absolute numbers, Figure 4 reveals that the fluctuations in both series happen at similar times.

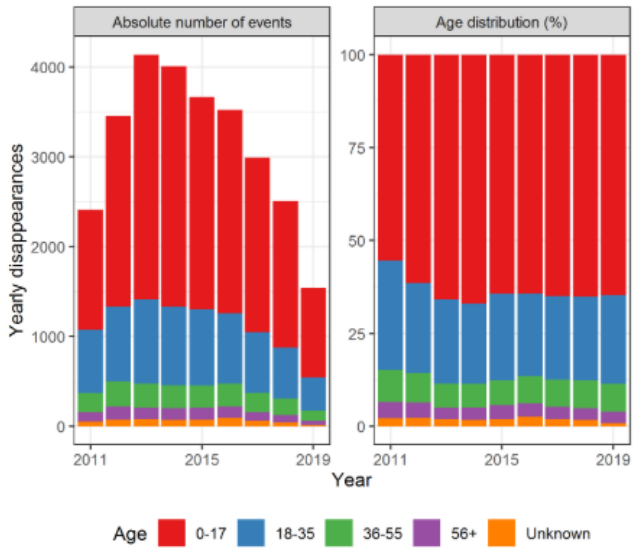

Figure 3: Yearly number of new disappearances by age group of the disappeared person and age composition of the population of disappeared persons in Guatemala. Data grouped by age comes from the Guatemalan National Police. Note: 2019 refers to data from January to August.

The monthly Twitter estimates, aggregated from individual-level Alerta Alba-Keneth records, also show the distribution of the disappeared by three age sub-groups intended to represent infants (0-4), children (4-12) and adolescents (13-17). We obtained this from the Twitter data since it is impossible to know the distribution of the population of youth by age or sex from the official statistics. Figure 5 reveals important differences within the age group 0-18. The number of disappeared aged 13-17 is higher than the younger age groups. Figure 5 also shows that women have been more likely to disappear according to the Twitter data, teenage girls in particular.

Anecdotal evidence suggests that many disappeared individuals may never be removed from the register of disappeared persons either because they are never found or because their re-appearance is not reported to the authorities. Parents of a disappeared child, for example, may fear losing custody after the child is found (Evelyn Espinoza, personal communication 19 September 2019). This results in a growing number of unresolved cases of disappeared youth. In Guatemala, these add up to more than 20,000 since 2010 alone. It is unclear how this growing backlog affects the production of national statistics.

4 NEXT STEPS

The methodology we propose in this work can be replicated to other Twitter accounts sharing information about disappeared people. We will extend our analysis to Brazil, where data at individual level is available. For for some states in Brazil we expect to have obtain individual-level information on disappeared persons which can then be directly compared to the Twitter data. By combining two data sources we will be able to do temporal and spatial analysis

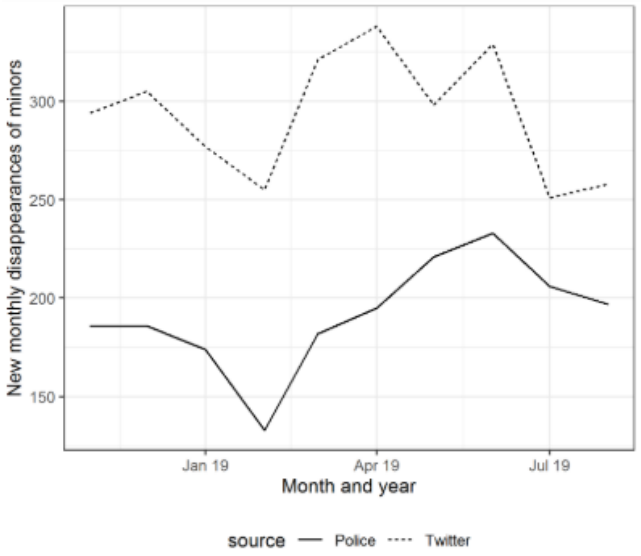

Figure 4: Monthly child and adolescent disappearances (under 18 years old): a comparative overview of data from the Guatemalan National Police and Alerta Alba-Keneth Twitter data.

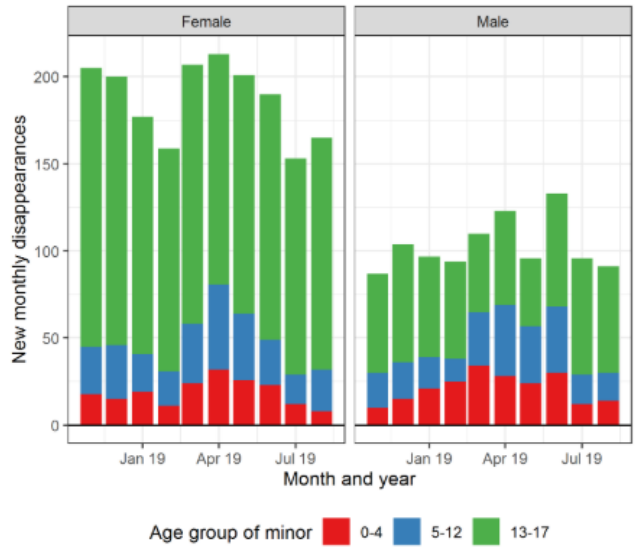

Figure 5: Age and sex distribution of the monthly disappeared youth (Oct 2018 to Aug 2019) according to the Alerta Alba-Keneth Twitter data.

looking at the data and place of disappearance. We are developing a real-time data collection platform to *nowcast* the demographic statistics of disappeared people using these data. Later on, we will use the skin color detection algorithm to estimate the composition of the population by ethnicity or race.

REFERENCES

García, J. and Aburto, J.M. (2019). The impact of violence on Venezuelan life expectancy and lifespan inequality. International journal of epidemiology 0(0).

Hendrickson, K. C. (2009) Morbidity, mortality, and parental grief: a review of the literature on the relationship between the death of a child and the subsequent health of parents. Palliat Support Care, 7(1):109–19

Li, J, Precht, D. H., Mortensen, P. B. and J. Olsen. (2003) Mortality in parents after the death of a child in Denmark: a nationwide follow-up study. Lancet, 361(9355): 363–7.

McIlwaine, C. and Moser, C.O.N. (2001) Violence and social capital in urban poor communities: perspectives from Colombia and Guatemala. Journal of International Development 13(7).

McCormick, T. H., Lee, H., Cesare, N., Shojaie, A., Spiro, E. S. (2017). Using Twitter for Demographic and Social Science Research: Tools for Data Collection and Processing. Sociological Methods & Research, 46(3), 390–421.

Meneghel, S.N. and Hirakata, V.N. (2011) Femicides: female homicide in Brazil. Rev. Saúde Pública 45(3).

Rostila, M. Saarela, J, and Kawachi I. (2012) Mortality in parents following the death of a child: a nationwide follow-up study from Sweden.Journal of Epidemiology and Community Health, 66(10):927–933.

Sharifara, A., Rahim, M. S. M., Anisi, Y. (2014, August). A general review of human face detection including a study of neural networks and Haar feature-based cascade classifier in face detection. In 2014 International Symposium on Biometrics and Security Technologies (ISBAST) (pp. 73-78). IEEE.

Soares-Filho, A.M. (2011) Homicide victimization according to racial characteristics in Brazil. Rev. Saúde Pública 45(4).

Tsoi, K. K., Chan, N. B., Chan, F. C., Zhang, L., Lee, A. C., Meng, H. M. (2018). How can we better use Twitter to find a person who got lost due to dementia?. npj Digital Medicine, 1(1), 14.

Telles, Edward, and Tianna Paschel. "Who Is Black, White, or Mixed Race? How Skin Color, Status, and Nation Shape Racial Classification in Latin America." American Journal of Sociology 120, no. 3 (2014): 864-907.