What Were You Thinking Nine Month Ago?
Using Twitter for Fertility Nowcasting Over Time and Space

Preliminary draft prepared for submission to PAA 2019

Dariya Ordanovich* [1, 2, 3], Diego Ramiro Fariñas[3], Francesco C. Billari[4], Antonia Tugores[5], Francisco Viciana[6], José J. Ramasco[5]

1. ESRI España, Madrid, Spain
2. Universidad Complutense de Madrid, Madrid, Spain
3. Spanish National Research Council, Madrid, Spain
4. Bocconi University, Milan, Italy
5. Instituto de Física Interdisciplinar y Sistemas Complejos IFISC (CSIC-UIB), Palma de Mallorca, Spain
6. Institute of Statistics and Cartography of Andalusia, Seville, Spain

A growing body of literature on methods and outcomes to leverage Twitter data in different domains reveals both large potentials and limitations of this innovative data source. However, these data are of an exceptional interest for demographers and health professionals as it brings in valuable real-time insights on demographic processes, mobility and human behavior, and which – as proved by many studies – frequently foresees official estimates. Getting comprehensive insights on shifts in fertility intentions beforehand through Twitter and hence understanding consequences for urban planning and changes in demand for services might a handy tool for researchers, city planners and policy makers. This study is built upon an extensive sample of over 100.000.000 tweets with geographic coordinates or location tags retrieved using real-time streaming API for the period from 2015 to 2018 in continental Spain. As a reference data we use official national and regional statistical datasets which are provided as aggregated time-series of births and deliveries and in the form of spatial grids with fertility estimates at minimal spatial resolution of 250 sq.m. in Andalusia. To identify focus group of Twitter users and analyze sentimental scores and latent semantic structures of their timelines we apply a wide range of machine learning techniques. Overall, this paper aims to show the potential of Twitter as a rich and timely data source valid for monitoring ('nowcasting') and anticipation of demographic trends, shifts and spatial distribution, introducing a significant added value to the official statistical production.

*Corresponding author: *dariya.ordanovich@esri.es*

Introduction

Fertility nowcasting, forecasting and its digital traces

Monitoring (or, in current terms 'nowcasting') and forecasting fertility is of paramount importance in several domains. Fertility is a key component of population projections, and it is therefore crucial to produce reliable estimations of its current levels (nowcasts), including for purposes of short-term predictions for both developing and affluent countries (Tomas Sobotka, 2004). Living in a society that changes more and more rapidly, we now require tools that are capable of monitoring these changes in real time, without the need to anticipate census operations or large surveys.

In general, and despite UN projections for the beginning of the XXI century (UN, 2001), the pace of fertility decline in more than 15 countries in sub-Saharan Africa, several countries in Asia and Latin America has slowed down, thus pointing out to the need of reconsideration of assumptions and sources included in the forecast, as estimating demographic indicators for developing countries is rather challenging in general terms (Alkema, Raftery, Gerland, Clark, & Pelletier, 2012). In developed countries, on the other hand, monitoring fertility, specifically its short-term fluctuations, becomes increasingly important, especially for measuring the effects of shifts such as the one provoked by the Great Recession (Matysiak, Sobotka, & Vignoli, 2018; Ramiro-Fariñas, Viciana-Fernandez, & Montañes Cobo, 2017; T. Sobotka, Skirbekk, & Philipov, 2011). Upon the whole, it is highlighted by multiple studies that the variations in fertility trends have direct and significant effects on the future size and age structure of the population (Bongaarts, 2008; Casterline, 2001).

At the global, regional and national level, population projections are built upon a variety of sources such as national censuses, vital registrations, immigration statistics and demographic surveys. New and innovative data sources, like web engine search queries or social media posts could complement existing practices and provide new insights into demographic behavior by relying on the 'digital breadcrumbs' that fertility decisions leave (Billari, D'Amuri, & Marcucci, 2016; Zagheni & Weber, 2015). These 'Big Data' therefore offer hints on short-term changes of fertility intentions, as well as on fertility trends and the geographical distribution of fertility. For this reason, 'nowcasting' - the prediction of the present using complementary information (R. Varian & Choi, 2009), introduces a significant added value to the official statistical production at a marginal cost, with the data produced by third parties (Struijs, Braaksma, & Daas, 2014).

Twitter as data source: justification of choice and limitations

Twitter is an online news and social networking service on which users post and interact with short messages known as "tweets". Twitter is somewhat a unique social media platform in terms of its

infrastructure, providing a large part of the generated data through an Application Programming Interface (API). Although Twitter is at lower levels with respect to Facebook and WhatsApp in terms of the total number of monthly active users, its data have the advantage of being available much more easily and in real time.

Along with other social media platforms, Twitter has shown a significant forecasting potential (Asur & Huberman, 2010): tweeting rates and sentiment polarity are commonly used for sales (Dijkman, Ipeirotis, Aertsen, & van Helden, 2015) or movie market revenues and rating predictions (Schmit & Wubben, 2015; Shim & Pourhomayoun, 2017). Network communication, influences on user opinion and information diffusion also appear within widely distributed research topics, especially in the context of electoral campaigns (Baviera 2018; El Bacha & Zin, 2019; Jungherr, 2016). It is worth mentioning the issue of Twitter data non-representativeness as a rather debated topic: whereas some studies find a particular skewness in user profiles towards more economically and educationally advantageous urban populations beneficial (e.g. ´elite´ behavior for political communication) (Blank, 2017), others maintain the discussion on its limitations and ambiguity for making valid inferences in other domains open (Mellon & Prosser, 2017; Soto et al., 2018; Tufekci, 2014). Another caveat for researchers might be hidden behind the peculiarities of how the location of Twitter messages is recorded: we discuss this matter in detail with regards to our specific study in the next section. Nevertheless, it has been proven that for certain aspects (e.g. population density estimation, mobility or urban land use) mobile phone records, surveys and geolocated Twitter data produce similar results at fine spatial scales. (Barlacchi et al., 2015; Patel et al., 2016).

Broadly, analytical insights using Twitter as the primary source for analyzing various aspects of population and demographic patterns can be grouped into following categories:

1. **Public health topics, including mental health and psychological wellbeing:**
    a. Monitoring the prevalence of unhealthy eating behavior (Abbar, Mejova, & Weber, 2015; Widener & Li, 2014), specifically regarding obesity or diabetes patterns (C. Nguyen et al., 2017; Ghosh & Guha, 2013; Harris, Moreland-Russell, Tabak, Ruhr, & Maier, 2014; Karami, Dahl, Turner-McGrievy, Kharrazi, & Shaw Jr, 2018) and physical activity (Nguyen et al., 2016; Zhang et al., 2013).
    b. Improving disease surveillance (Gomide et al., 2011; K. Lee, Agrawal, & Choudhary, 2013), tracking the transmission (Sadilek, Kautz, & Silenzio, 2012) and prevalence of diseases, with special attention given to predicting trends in seasonal influenza (Achrekar, Gandhe, Lazarus, Ssu-Hsin, & Liu, 2011; Broniatowski, Paul, & Dredze, 2013; Li & Cardie, 2013; Nagar et al., 2014).

c. Exploring trends in tobacco, alcohol and drug use, most commonly included as part of a wider analysis aiming to build a comprehensive picture of spatial differentiation at a neighborhood level (Meng, Kath, Li, & Nguyen, 2017; Nguyen et al., 2017).
d. Detecting signs of depression, self-harm, and suicidality (Cavazos-Rehg et al., 2016; Yang & Mu, 2015).

2. **Comprehensive analysis of neighborhoods, event detection and social response:**
   a. General classification of space and environmental exposure (Hansen Andrew et al., 2013); characterization of urban areas (R. Lee, Wakamiya, & Sumiya, 2013; Wakamiya, Lee, & Sumiya, 2011).
   b. Event detection (C.-H. Lee, Yang, Chien, & Wen, 2011), estimation of crowd sizes (Botta, Moat, & Preis, 2015), geographical models of behavior (R. Lee & Sumiya, 2010b; R. Lee, Wakamiya, & Sumiya, 2011) and crime analysis (Ristea, Andresen, & Leitner, 2018).
3. **Stratification of Twitter population by socio-demographic characteristics and analysis of demographic processes:**
   a. Profiling the Twitter population (e.g. predicting age, gender and ethnicity) (Culotta, Kumar, & Cutler, 2015; Fink, Kopecky, & Morawski, 2012; Mislove, Lehmann, Ahn, Onnela, & Rosenquist, 2011; Luke Sloan, 2017; Wood-Doughty, Andrews, Marvin, & Dredze, 2018; Yildiz, Munson, Vitali, Tinati, & Holland, 2017).
   b. Chronic stress expressed in Twitter language and cardiovascular mortality (J. Eichstaedt et al., 2018; J. C. Eichstaedt et al., 2015).
   c. Human mobility at national and cross-border scales (Blanford, Huang, Savelyev, & MacEachren, 2015; Jurdak et al., 2015; Zagheni & Weber, 2015).
   d. Attitudes on fertility and parenthood (Sulis et al., 2016).
   e. Connecting demographic behavior and trends with "soft" measures: complementary information on attitudes, values, feelings or intentions (Mencarini, 2018; Sulis, Lai, Vinai, & Sanguinetti, 2015).

The last two aspects have not yet been extensively studied. In this paper, we focus on the development of both a conceptual framework and a methodological workflow that allows to comprehensively analyze fertility fluctuations in time and space using Twitter data.

Objective

This paper evaluates the usefulness of Twitter big data for fertility research and explores potentials and limitations for leveraging these data. More precisely, we aim to utilize geolocated[1] and geotagged[2] Twitter data as a way of nowcasting fertility intentions and short-term fertility changes in time and space.

[1] Containing precise coordinates (latitude and longitude), registered and emitted by device used for tweeting
[2] Marked with a location name chosen by user from a drop-down list of 20 closest locations at the moment of posting

Data and Methods

Primary source

This study is built upon an extensive Twitter sample retrieved using real-time streaming API which covers four years of user activities (2015-2018), and either contains precise tweets' coordinates (i.e. longitude and latitude) or information on tweets' location deliberately added by users. This sample represents over 100 million of tweets and is limited by bounding box (extreme coordinates) of continental Spain.

Geocoding tweets

A plethora of studies highlights the importance of geolocated Twitter data, and suggests that these data can substantially improve our understanding of continuous human mobility (Hawelka et al., 2014; Zagheni, Garimella, Weber, & State, 2014), social interactions (Grabowicz, Ramasco, Gonçalves, & Eguíluz, 2014; Osman, Ahmad, Halizah, & Basiron, 2017) and behaviors in response to certain events (C.-H. Lee et al., 2011; R. Lee & Sumiya, 2010a). However, since April 2015 Twitter collects, stores and uses the precise location only for tweets posted by users who have **intentionally** enabled the option to *Share Precise Location* feature on their devices. Once this feature is enabled, users posting via official Twitter application can optionally attach a location from a dropdown list (such as a city, neighborhood or a specific name of a place i.e. point of interest) of their choice to a tweet. Tweets geotagged by users and posted prior to aforementioned date by default include both a location label and device's precise location (Kinder-Kurlanda, Weller, Zenk-Möltgen, Pfeffer, & Morstatter, 2017). As the privacy implications of geocoding became more transparent to Twitter users in April 2015, this resulted in a sharp decrease in the number of posts with precise location, and by the beginning of 2018 geo-tagged tweets represented only about 2% of all tweets and 3% of Twitter users (Burton, Tanner, Giraud-Carrier, West, & Barnes, 2012), thus severely limiting the potential of large-scale analyses. Nevertheless, if users who do not opt for enabling coordinates-sharing choose instead to tag their tweets with a particular location from a drop-down list with 20 closest locations – and if this tag is specific enough – it is then possible to infer the coordinates from the location name. In particular, this study aims to maximize the geographic component of analysis by inferring spatial location associated with a Twitter message based on *1.* location tags via StreetMap Premium geolocator[3] and *2.* information contained in proper Twitter messages and user profiles. In posterior steps of analysis we take into consideration potential differences in demographic composition of users who use coordinates-sharing and those who opt for geotagging (L. Sloan & Morgan, 2015).

[3] StreetMap Premium is based on commercial street reference data from leading global and local street data suppliers and is provided by Esri Spain under correspondent licensing conditions.

Annotation and sentiment analysis

After removing accounts that are classified as unsuitable for the analysis, spammy or automated, the text in each tweet is cleaned and annotated based on the natural language processing toolkit which contains pre-trained Universal Dependency models for Spanish and English languages (Wijffels, Straka, & Straková, 2018). The annotated dataset enables us to identify different parts of speech[4] and work with lemmas (instead of stems[5]). We then assign sentimental score to every term in a tweet based on the available for academic research Spanish and English lexicons and calculate overall tweet polarity.

Focus group identification

To identify the focus group (i.e. group of users who have posted tweet/tweets of direct relevance to childbirth and/or pregnancy topics at any point of time during the study period and thus are considered possessing a direct interest in these topics) we filter the entire dataset by key lemmas (e.g. "embarazo", "pregnant" and so on, slightly expanding the dictionary suggested by (Sulis et al., 2016)) and then classify a small random sub-sample within this selection into the categories listed in the ***Table 1*** in order to create a corpus on pregnancy and childbirth.

Table 1 Definitions used for manual sample classification

	Relevance	Description
Class 1	Direct	Individuals tweeting about own pregnancy/childbirth;
Class 2	Indirect	Individuals tweeting about pregnancy/childbirth of people they know, a friend or family member, *or*
Class 3		Individuals tweeting about the topic without indication of direct involvement in the process of either pregnancy or childbirth; could be politics, expression of opinion on the news, *or*
Class 4		Professionals tweeting about the topic e.g. photographers, product sales and so on;
Class 0	Off-topic	Complete irrelevance, noise.

Upon construction of such corpus, we automate and generalize the identification of focus group in the entire sample by using machine learning techniques. Some features of this dataset (e.g. its intrinsically

[4] Here we give our preference to primarily nouns and adjectives, however for descriptive frequency and co-occurrence analysis we also include verbs.

[5] Lemmatization (the process of converting a given word to the base form of all its inflectional forms) considers the morphological analysis of the words, whereas stemming algorithms work by removing the end (and sometimes the beginning) of a given word.

unbalanced structure with small subset of words naturally appearing more frequently in the text with respect to other words) imply that several balancing techniques (e.g. Term Frequency-Inverse Document Frequency, or *tf-idf* as commonly referred to in text analytics) are to be used to increase the predictive performance of the global method.

In the following step, we test several algorithms most commonly used for supervised text classification (including Support Vector Machines (Basu, Walters, & Shepherd, 2003) and Neural Networks (Patil, Gune, & Nene, 2017)), and implement the most suitable one to create a model with appropriate level of accuracy. To assess the predictive performance of the models, we analyze confusion matrices and calibrate the model until the highest level of specificity is reached.

Finally, in the last step, we use this trained and tuned model to predict classes on the rest of the filtered by key lemmas dataset.

Topic modelling

We narrow down our sample by selecting entire tweeting timeline for each user in the focus group, leveraging this selection to discover its latent semantic structures. To do so, we test various steps for an adequate temporal aggregation and evaluate different unsupervised classification machine learning methods suitable for topic modelling (e.g. Latent Dirichlet Allocation technique (Welbers, Van Atteveldt, & Benoit, 2017)). Additionally, we conduct a descriptive frequency analysis of terms and hashtags with the selected time step using a naïve Bayes approach (also known as a bag-of-words approach) (Proksch & Slapin, 2009).

Spatial analysis

Furthermore, we perform a series of spatial operations to analyze distributions and clusters in the dataset corresponding to the focus group, testing it for dependencies with the ground truth measurements.

In the first instance, we target to identify locations of the most frequent activity (i.e. presumably home and work places) for users in the focus group. For each user, we identify density clusters for the entire period of tweeting timeline, which we then classify into home and work locations based on the cluster size and most prevalent tweeting time.

Using the Getis-Ord Gi* statistic we identify statistically significant hot and cold spots for the clusters of tweets associated with home user locations across Spain. We also perform an emerging hot spot analysis to identify trends in the clustering of point densities in a so-called space-time cube.

Possessing data on fertility estimates at the 250 sq.m. grid level in Andalusia, we first aim to analyze the relationship between density of tweets in a focus group within these geographical units and fertility outcomes. On the further steps, we scale down this analysis to the grid level of 10 sq.km. with national coverage.

Ground truth data

The reference data on population used in this study are retrieved from official statistical sources and contain:

1. Aggregated to regional (autonomous community of Andalusia) level daily counts of deliveries (2012-2017)[6]
2. Spatial grid with fertility estimates (Total Fertility Rates and Standardized Fertility Rates (SFR) [7]) at 250 sq.m. and 1 sq.km. resolution
3. Aggregated at national level daily counts of births (2012-2016)[8]
4. Spatial grid with fertility estimates at 10 sq.km. resolution[9]

Preliminary results

In this section we present some of the preliminary results obtained when working with Twitter and Ground Truth data for the region of Andalusia and period of 2015-2016 only. As mentioned previously, the objective of this study is to perform the analysis at the national scale with an extended temporal coverage.

General trends

After applying rolling average smoothing techniques on both national daily births (**Error! Reference source not found.**)/regional daily deliveries(**Figure 2**) counts and tweets of the focus group in Andalusia, we observed similar pattern in both time-series within a given time period.

---

[6] Source: Longitudinal Population Register of Andalusia, Institute of Statistics and Cartography of Andalusia. URL https://www.juntadeandalucia.es/institutodeestadisticaycartografia/fecundidad/index-en.htm

[7] SFR smoothed indicators for the first birth are calculated per cell (250 sq.m and 1 sq.km) of residence during the follow-up period of 12 years since 2001 and represent unique fertility maps stratified by sex and observation period. To reduce the variability of the indicator for small populations local Bayesian smoothing technique is applied. This method not only uses the information from the cell itself, but also considers cell "environment" to obtain a larger population and consider its influence on the fertility of a given cell. Based on bootstrap techniques a smoothed indicator with credibility intervals at an established confidence level is obtained, and classification of cells into 5 fertility groups is defined: low, moderately low, similar to the Andalusian average, moderately high and high. URL https://www.juntadeandalucia.es/institutodeestadisticaycartografia/fecundidad/metodologia/metodologia_mapa.pdf

[8] Source: Spanish National Institute of Statistics URL http://www.ine.es/welcome.shtml

[9] Currently in production.

Figure 1 Tweeting behavior in the focus group versus daily birth counts at the national level

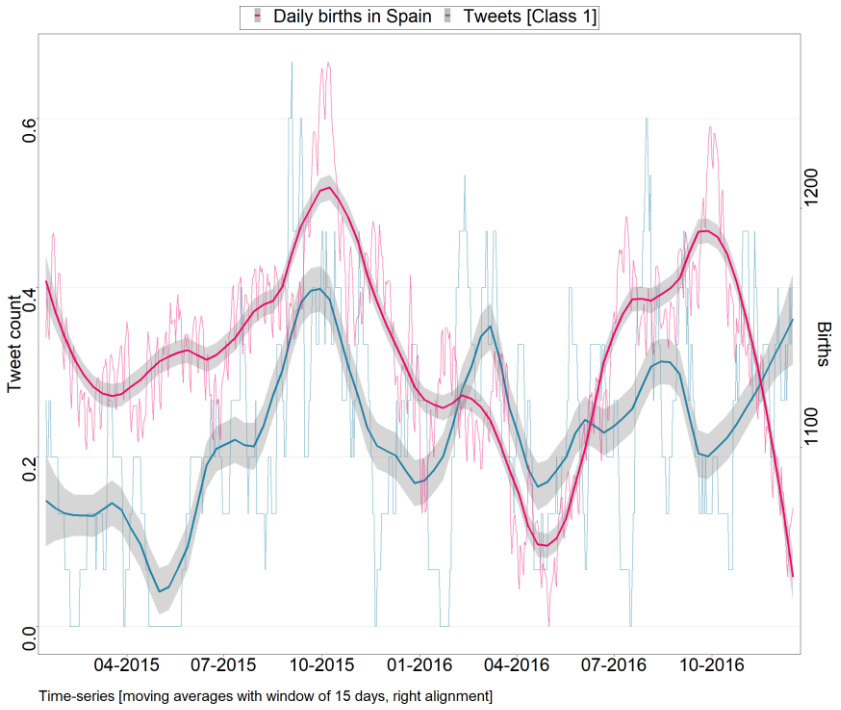
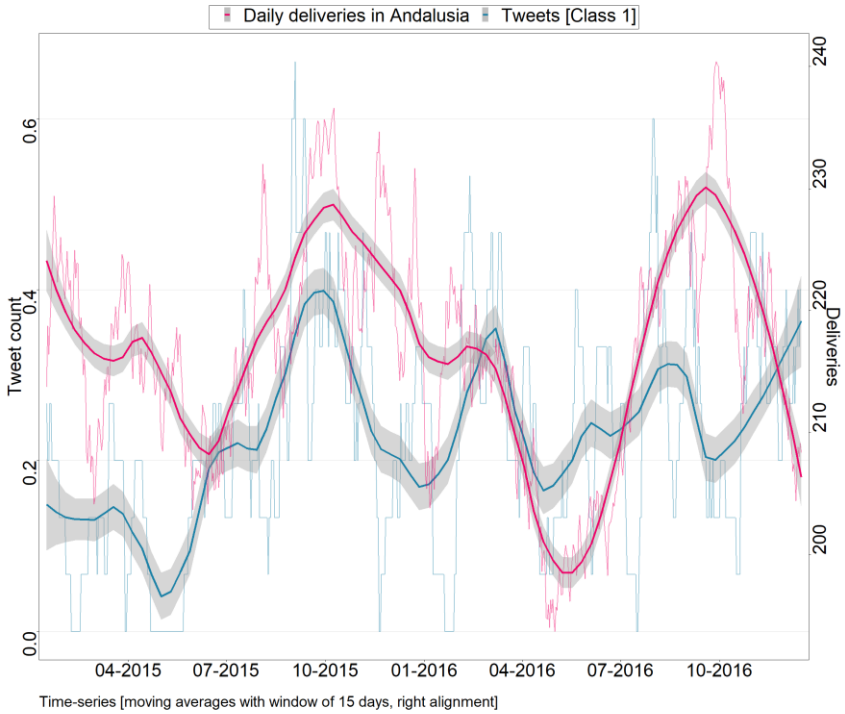


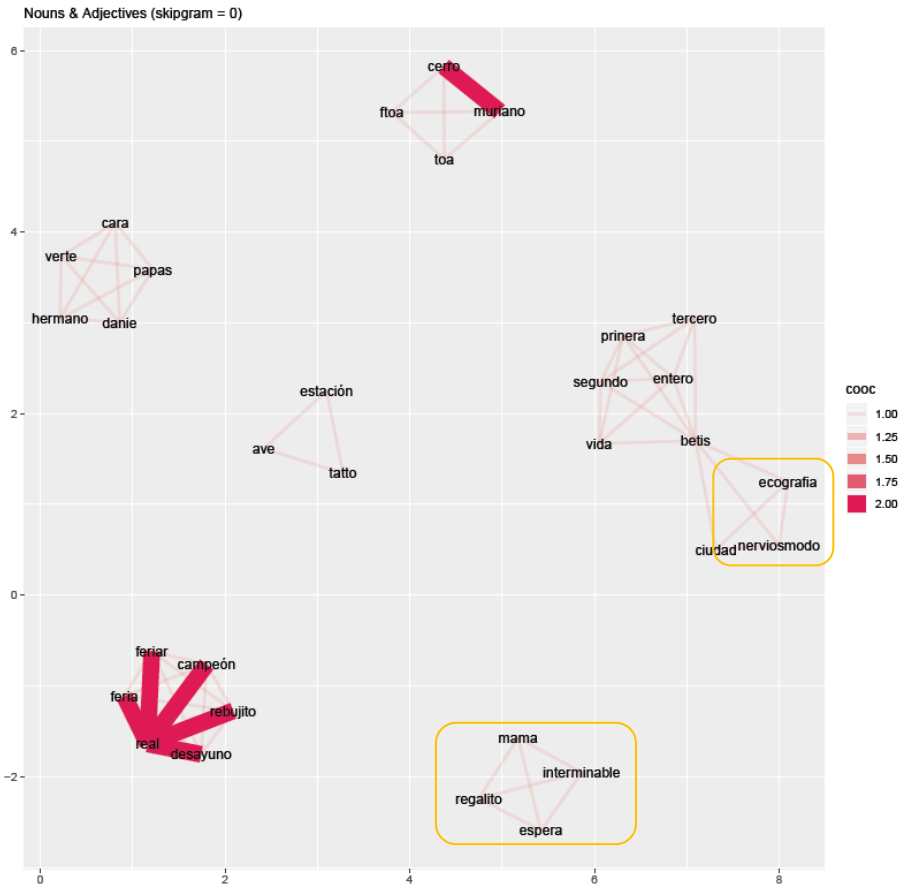
Figure 2 Tweeting behavior in the focus group versus daily delivery counts at the regional level



Term co-occurrence

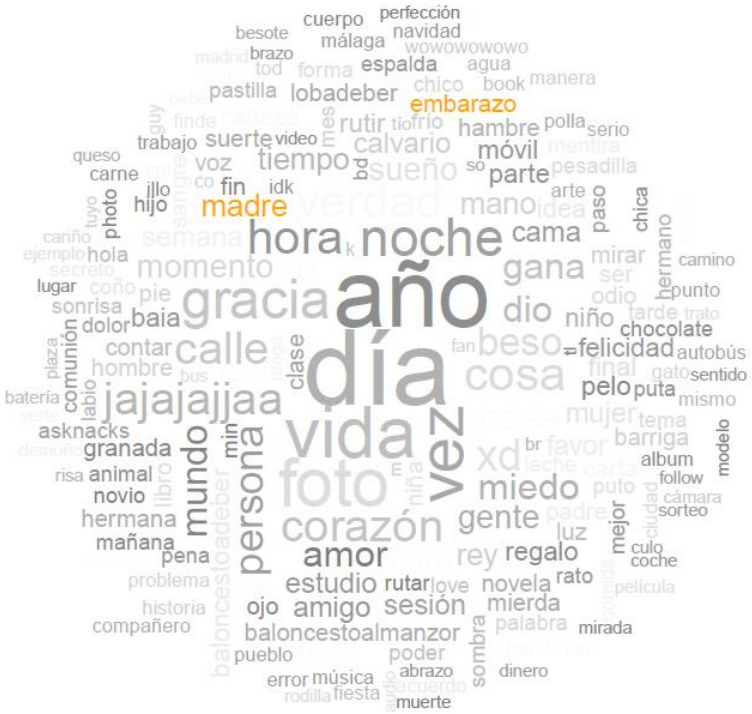
As part of the exploratory analysis, we evaluated co-occurrences between different parts of speech in the text. The graph displayed in **Figure 3** shows aggregation of consequent nouns and adjectives into distinct thematic clusters in a single user profile with several clusters indicating pregnancy-related topics (highlighted in yellow).

Figure 3 Co-occurrences of lemmas in a single user profile (aggregated for one month)



However, if aggregated to a wider time range (e.g. several months as shown in **Figure 4**) patterns become less visible and overgeneralized.

Figure 4 Word cloud of the most common lemmas appearing in a single user profile (aggregation for 3 months)

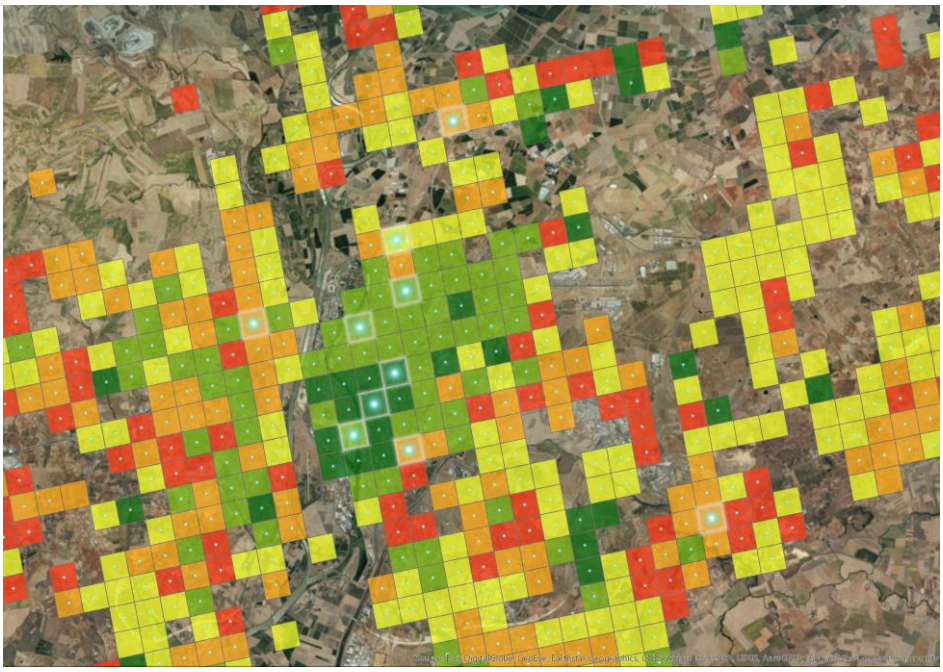


Spatial patterns

Exploratory maps for the urban areas of Seville (**Figure 5a**) and Málaga (**Figure 5b**) reveal distinct patterns. Whereas in Málaga tweeting rates of the focus group are higher in the periphery of the city and tend to coincide with areas of more intense SFR, Seville – the largest tourist hub of the province – sees a concentration with higher tweeting rate in the focus group within the city center characterized by low indicators of SFR.

Figure 5 Tweeting rates in focus group versus fertility intensity in Andalusia

a.



b.

Tweeting rate (focus group users in a cell/total Twitter users in a cell)

Classes of standardized fertility rates

Summary and future steps

Analytical insights using Twitter as the primary source are becoming increasingly popular for the analysis of various aspects in different domains, including analysis of population and demographic patterns. This paper targets to evaluate suitability of non-traditional data sources like Twitter for studying trends and patterns of fertility on the example of Spain. We use Twitter data with different geographic precision as a way of nowcasting fertility intentions and short-term changes in time and space. We enrich the analysis by testing the results of text mining against ground truth data provided by the Spanish National Institute of Statistics and the Institute of Statistics and Cartography of Andalusia. Overall, this paper aims to show the potential of Twitter as a rich and timely data source valid for the monitoring ('nowcasting'), anticipation of demographic trends, shifts and spatial distribution, introducing a significant added value to the official statistical production.

References

Abbar, S., Mejova, Y., & Weber, I. (2015). *You Tweet What You Eat: Studying Food Consumption Through Twitter*. Paper presented at the Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems, Seoul, Republic of Korea.

Achrekar, H., Gandhe, A., Lazarus, R., Ssu-Hsin, Y., & Liu, B. (2011, 10-15 April 2011). *Predicting Flu Trends using Twitter data*. Paper presented at the 2011 IEEE Conference on Computer Communications Workshops (INFOCOM WKSHPS).

Alkema, L., Raftery, A. E., Gerland, P., Clark, S. J., & Pelletier, F. (2012). Estimating trends in the total fertility rate with uncertainty using imperfect data: Examples from West Africa. *Demographic Research, 26*(15), 331-362.

Asur, S., & Huberman, B. A. (2010, 31 Aug.-3 Sept. 2010). *Predicting the Future with Social Media*. Paper presented at the 2010 IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology.

Barlacchi, G., De Nadai, M., Larcher, R., Casella, A., Chitic, C., Torrisi, G., . . . Lepri, B. (2015). A multi-source dataset of urban life in the city of Milan and the Province of Trentino. *Scientific Data, 2*, 150055. doi:10.1038/sdata.2015.55

Basu, A., Walters, C., & Shepherd, M. (2003, 6-9 Jan. 2003). *Support vector machines for text categorization*. Paper presented at the 36th Annual Hawaii International Conference on System Sciences, 2003. Proceedings of the.

Baviera , T. (2018). Influence in the political Twitter sphere: Authority and retransmission in the 2015 and 2016 Spanish General Elections. *European Journal of Communication, 33*(3), 321-337. doi:10.1177/0267323118763910

Billari, F., D'Amuri, F., & Marcucci, J. (2016). *Forecasting Births Using Google*.

Blanford, J. I., Huang, Z., Savelyev, A., & MacEachren, A. M. (2015). Geo-Located Tweets. Enhancing Mobility Maps and Capturing Cross-Border Movement. *PLOS ONE, 10*(6), e0129202. doi:10.1371/journal.pone.0129202

Blank, G. (2017). The Digital Divide Among Twitter Users and Its Implications for Social Research. *Social Science Computer Review, 35*(6), 679-697. doi:10.1177/0894439316671698

Bongaarts, J. (2008). Fertility Transitions in Developing Countries: Progress or Stagnation? *Studies in Family Planning, 39*(2), 105-110. doi:doi:10.1111/j.1728-4465.2008.00157.x

Botta, F., Moat, H. S., & Preis, T. (2015). Quantifying crowd size with mobile phone and Twitter data. *Royal Society Open Science, 2*(5). doi:10.1098/rsos.150162

Broniatowski, D. A., Paul, M. J., & Dredze, M. (2013). National and Local Influenza Surveillance through Twitter: An Analysis of the 2012-2013 Influenza Epidemic. *PLOS ONE, 8*(12), e83672. doi:10.1371/journal.pone.0083672

Burton, S. H., Tanner, K. W., Giraud-Carrier, C. G., West, J. H., & Barnes, M. D. (2012). "Right Time, Right Place" Health Communication on Twitter: Value and Accuracy of Location Information. *J Med Internet Res, 14*(6), e156. doi:10.2196/jmir.2121

C. Nguyen, Q., D. Brunisholz, K., Yu, W., McCullough, M., Hanson, H., Litchman, M., . . . Smith, K. (2017). *Twitter-derived neighborhood characteristics associated with obesity and diabetes* (Vol. 7).

Casterline, J. B. (2001). The Pace of Fertility Transition: National Patterns in the Second Half of the Twentieth Century. *Population and Development Review, 27*, 17-52.

Cavazos-Rehg, P. A., Krauss, M. J., Sowles, S., Connolly, S., Rosas, C., Bharadwaj, M., & Bierut, L. J. (2016). A content analysis of depression-related Tweets. *Comput Human Behav, 54*, 351-357. doi:10.1016/j.chb.2015.08.023

Culotta, A., Kumar, N. R., & Cutler, J. (2015). *Predicting the Demographics of Twitter Users from Website Traffic Data*.

Dijkman, R., Ipeirotis, P., Aertsen, F., & van Helden, R. (2015). *Using Twitter to Predict Sales: A Case Study*.

Eichstaedt, J., Schwartz, H., Giorgi, S., Kern, M., Park, G., Sap, M., . . . Ungar, L. (2018). *More Evidence that Twitter Language Predicts Heart Disease: A Response and Replication*.

Eichstaedt, J. C., Schwartz, H. A., Kern, M. L., Park, G., Labarthe, D. R., Merchant, R. M., . . . Seligman, M. E. (2015). Psychological language on Twitter predicts county-level heart disease mortality. *Psychol Sci, 26*(2), 159-169. doi:10.1177/0956797614557867

El Bacha, R., & Zin, T. T. (2019). *A Survey on Influence and Information Diffusion in Twitter Using Big Data Analytics*.

Fink, C., Kopecky, J., & Morawski, M. (2012). *Inferring Gender from the Content of Tweets: A Region Specific Example*.

Ghosh, D. D., & Guha, R. (2013). What are we 'tweeting' about obesity? Mapping tweets with Topic Modeling and Geographic Information System. *Cartogr Geogr Inf Sci, 40*(2), 90-102. doi:10.1080/15230406.2013.776210

Gomide, J., Veloso, A., Wagner Meira, J., Almeida, V., Benevenuto, F., Ferraz, F., & Teixeira, M. (2011). *Dengue surveillance based on a computational model of spatio-temporal locality of Twitter*. Paper presented at the Proceedings of the 3rd International Web Science Conference, Koblenz, Germany.

Grabowicz, P. A., Ramasco, J. J., Gonçalves, B., & Eguíluz, V. M. (2014). Entangling Mobility and Interactions in Social Media. *PLOS ONE, 9*(3), e92196. doi:10.1371/journal.pone.0092196

Hansen Andrew, S., Johannes, C. E., Margaret, L. K., Lukasz, D., Richard, E. L., Megha, A., . . . Lyle, U. (2013). Characterizing Geographic Variation in Well-Being Using Tweets. *International AAAI Conference on Web and Social Media; Seventh International AAAI Conference on Weblogs and Social Media*.

Harris, J. K., Moreland-Russell, S., Tabak, R. G., Ruhr, L. R., & Maier, R. C. (2014). Communication about childhood obesity on Twitter. *American journal of public health, 104*(7), e62-69. doi:10.2105/AJPH.2013.301860

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. doi:10.1080/15230406.2014.890072

Jungherr, A. (2016). Twitter use in election campaigns: A systematic literature review. *Journal of Information Technology & Politics, 13*(1), 72-91. doi:10.1080/19331681.2015.1132401

Jurdak, R., Zhao, K., Liu, J., AbouJaoude, M., Cameron, M., & Newth, D. (2015). Understanding Human Mobility from Twitter. *PLOS ONE, 10*(7), e0131469. doi:10.1371/journal.pone.0131469

Karami, A., Dahl, A., Turner-McGrievy, G., Kharrazi, H., & Shaw Jr, G. (2018). *Characterizing Diabetes, Diet, Exercise, and Obesity Comments on Twitter* (Vol. 38).

Kinder-Kurlanda, K., Weller, K., Zenk-Möltgen, W., Pfeffer, J., & Morstatter, F. (2017). Archiving information from geotagged tweets to promote reproducibility and comparability in social media research. *Big Data & Society, 4*(2), 2053951717736336. doi:10.1177/2053951717736336

Lee, C.-H., Yang, H.-C., Chien, T.-F., & Wen, W.-S. (2011). *A Novel Approach for Event Detection by Mining Spatio-temporal Information on Microblogs*. Paper presented at the Proceedings of the 2011 International Conference on Advances in Social Networks Analysis and Mining.

Lee, K., Agrawal, A., & Choudhary, A. (2013). *Real-time disease surveillance using Twitter data: demonstration on flu and cancer*. Paper presented at the Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining, Chicago, Illinois, USA.

Lee, R., & Sumiya, K. (2010a). *Measuring geographical regularities of crowd behaviors for Twitter-based geo-social event detection*.

Lee, R., & Sumiya, K. (2010b). *Measuring geographical regularities of crowd behaviors for Twitter-based geo-social event detection*. Paper presented at the Proceedings of the 2nd ACM SIGSPATIAL International Workshop on Location Based Social Networks, San Jose, California.

Lee, R., Wakamiya, S., & Sumiya, K. (2011). Discovery of unusual regional social activities using geo-tagged microblogs. *World Wide Web, 14*(4), 321-349. doi:10.1007/s11280-011-0120-x

Lee, R., Wakamiya, S., & Sumiya, K. (2013). Urban area characterization based on crowd behavioral lifelogs over Twitter. *Personal and Ubiquitous Computing, 17*(4), 605-620. doi:10.1007/s00779-012-0510-9

Li, J., & Cardie, C. (2013). *Early Stage Influenza Detection from Twitter*.

Matysiak, A., Sobotka, T., & Vignoli, D. (2018). The great recession and fertility in Europe: a sub-national analysis. *Vienna Institute of demography*.

Mellon, J., & Prosser, C. (2017). Twitter and Facebook are not representative of the general population: Political attitudes and demographics of British social media users. *Research & Politics, 4*(3), 2053168017720008. doi:10.1177/2053168017720008

Mencarini, L. (2018). *The Potential of the Computational Linguistic Analysis of Social Media for Population Studies*.

Meng, H.-W., Kath, S., Li, D., & Nguyen, Q. C. (2017). National substance use patterns on Twitter. *PLOS ONE, 12*(11), e0187691. doi:10.1371/journal.pone.0187691

Mislove, A., Lehmann, S., Ahn, Y.-Y., Onnela, J.-P., & Rosenquist, J. N. (2011). *Understanding the Demographics of Twitter Users*.

Nagar, R., Yuan, Q., Freifeld, C. C., Santillana, M., Nojima, A., Chunara, R., & Brownstein, J. S. (2014). A case study of the New York City 2012-2013 influenza season with daily geocoded Twitter data from temporal and spatiotemporal perspectives. *J Med Internet Res, 16*(10), e236. doi:10.2196/jmir.3416

Nguyen, Q. C., Li, D., Meng, H. W., Kath, S., Nsoesie, E., Li, F., & Wen, M. (2016). Building a National Neighborhood Dataset From Geotagged Twitter Data for Indicators of Happiness, Diet, and Physical Activity. *JMIR Public Health Surveill, 2*(2), e158. doi:10.2196/publichealth.5869

Nguyen, Q. C., McCullough, M., Meng, H. W., Paul, D., Li, D., Kath, S., . . . Li, F. (2017). Geotagged US Tweets as Predictors of County-Level Health Outcomes, 2015-2016. *American journal of public health, 107*(11), 1776-1782. doi:10.2105/ajph.2017.303993

Osman, A. N. S., Ahmad, S. S. S., Halizah, & Basiron. (2017). *Impact of Twitter on Human Interaction*.

Patel, N., Stevens, F., Huang, Z., Gaughan, A., Elyazar, I., & Tatem, A. (2016). *Improving Large Area Population Mapping Using Geotweet Densities: Improving Large Area Population Mapping Using Geotweet Densities* (Vol. 21).

Patil, S., Gune, A., & Nene, M. (2017, 1-2 Aug. 2017). *Convolutional neural networks for text categorization with latent semantic analysis*. Paper presented at the 2017 International Conference on Energy, Communication, Data Analytics and Soft Computing (ICECDS).

Proksch, S.-O., & Slapin, J. B. (2009). How to Avoid Pitfalls in Statistical Analysis of Political Texts: The Case of Germany. *German Politics, 18*(3), 323-344. doi:10.1080/09644000903055799

R. Varian, H., & Choi, H. (2009). *Predicting the Present with Google Trends* (Vol. 88).

Ramiro-Fariñas, D., Viciana-Fernandez, F., & Montañes Cobo, V. (2017). Will highly educated women have more children in the future? In Southern Europe, it will largely depend on labour market conditions. In *Vienna Yearbook of Population Research* (Vol. 15, pp. 49–54).

Ristea, A., Andresen, M. A., & Leitner, M. (2018). Using tweets to understand changes in the spatial crime distribution for hockey events in Vancouver. *The Canadian Geographer / Le Géographe canadien, 62*(3), 338-351. doi:doi:10.1111/cag.12463

Sadilek, A., Kautz, H., & Silenzio, V. (2012). *Predicting disease transmission from geo-tagged micro-blog data*. Paper presented at the Proceedings of the Twenty-Sixth AAAI Conference on Artificial Intelligence, Toronto, Ontario, Canada.

Schmit, W., & Wubben, S. (2015). *Predicting Ratings for New Movie Releases from Twitter Content*.

Shim, S., & Pourhomayoun, M. (2017). *Predicting Movie Market Revenue Using Social Media Data*.

Sloan, L. (2017). Who Tweets in the United Kingdom? Profiling the Twitter Population Using the British Social Attitudes Survey 2015. *Social Media + Society, 3*(1), 2056305117698981. doi:10.1177/2056305117698981

Sloan, L., & Morgan, J. (2015). Who Tweets with Their Location? Understanding the Relationship between Demographic Characteristics and the Use of Geoservices and Geotagging on Twitter. *PLOS ONE, 10*(11), e0142209. doi:10.1371/journal.pone.0142209

Sobotka, T. (2004). *Postponement of Childbearing and Low Fertility in Europe*.

Sobotka, T., Skirbekk, V., & Philipov, D. (2011). Economic recession and fertility in the developed world. *Popul Dev Rev, 37*(2), 267-306.

Soto, A., Ryan, C., Peña Silva, F., Das, T., Wolkowicz, J., Milios, E., & Brooks, S. (2018). *Data Quality Challenges in Twitter Content Analysis for Informing Policy Making in Health Care*.

Struijs, P., Braaksma, B., & Daas, P. J. (2014). Official statistics and Big Data. *Big Data & Society, 1*(1), 2053951714538417. doi:10.1177/2053951714538417

Sulis, E., Bosco, C., Patti, V., Lai, M., Farías, D. I. H., Mencarini, L., . . . Vignoli, D. (2016). *Subjective Well-Being and Social Media. A Semantically Annotated Twitter Corpus on Fertility and Parenthood*.

Sulis, E., Lai, M., Vinai, M., & Sanguinetti, M. (2015). *Exploring sentiment in social media and official statistics: A general framework* (Vol. 1351).

Tufekci, Z. (2014). *Big Questions for Social Media Big Data: Representativeness, Validity and Other Methodological Pitfalls*.

UN. (2001). *World Population Prospects: The 2000 Revision*. Retrieved from New York, NY 10017: http://www.un.org/esa/population/publications/wpp2000/highlights.pdf

Wakamiya, S., Lee, R., & Sumiya, K. (2011, 2011//). *Urban Area Characterization Based on Semantics of Crowd Activities in Twitter*. Paper presented at the GeoSpatial Semantics, Berlin, Heidelberg.

Welbers, K., Van Atteveldt, W., & Benoit, K. (2017). Text Analysis in R. *Communication Methods and Measures, 11*(4), 245-265. doi:10.1080/19312458.2017.1387238

Widener, M. J., & Li, W. (2014). Using geolocated Twitter data to monitor the prevalence of healthy and unhealthy food references across the US. *Applied Geography, 54*, 189-197. doi:https://doi.org/10.1016/j.apgeog.2014.07.017

Wijffels, J., Straka, M., & Straková, J. (2018). udpipe: Tokenization, Parts of Speech Tagging, Lemmatization and Dependency Parsing with the 'UDPipe' 'NLP' Toolkit (Version 0.7). CRAN. Retrieved from https://CRAN.R-project.org/package=udpipe

Wood-Doughty, Z., Andrews, N., Marvin, R., & Dredze, M. (2018). *Predicting Twitter User Demographics from Names Alone*.

Yang, W., & Mu, L. (2015). *GIS analysis of depression among Twitter users* (Vol. 60).

Yildiz, D., Munson, J., Vitali, A., Tinati, R., & Holland, J. A. (2017). Using Twitter data for demographic research. *Demographic Research, 37*(46), 1477-1514.

Zagheni, E., Garimella, V. R. K., Weber, I., & State, B. (2014). *Inferring international and internal migration patterns from Twitter data*. Paper presented at the Proceedings of the 23rd International Conference on World Wide Web, Seoul, Korea.

Zagheni, E., & Weber, I. (2015). Demographic research with non-representative internet data. *International Journal of Manpower, 36*(1), 13-25. doi:doi:10.1108/IJM-12-2014-0261

Zhang, N., Campo, S., Janz, K., Eckler, P., Yang, J., G Snetselaar, L., & Signorini, A. (2013). *Electronic Word of Mouth on Twitter About Physical Activity in the United States: Exploratory Infodemiology Study* (Vol. 15).