3rd International Conference on Advanced Research Methods and Analytics (CARMA2020)
Universitat Politècnica de València, València, 2020
DOI: http://dx.doi.org/10.4995/CARMA2020.2020.11359

Citizens' attention in Madrid City through the study of personalized records

Pilar Rey del Castillo
Instituto de Estudios Fiscales, Ministry of Finance, Spain.

Abstract

The datification of our daily lives in the Big Data era is producing a huge amount of information about processes and activities that were previously invisible or at least difficult to grasp, leading to new opportunities and challenges for analysis.

Examples of some data available are the tens of million of Personalized Attention Records that can be downloaded from the open data portal offered by the local government of Madrid City. These records become a sort of counterpart from the call receiver's perspective of the Call Detail Records produced by telecom providers. They are stored as a result of a front office tool retaining some information from a range of different communication channels to manage the interaction with users.

The paper explores the data contained on these Personalized Attention Records to help improve customer attention services. It emphasizes the study of the topics that concern the citizens and the different channels dealing with the services, using Natural Language Processing and other tools.

Keywords: *Big Data; Call Records; Natural Language Processing.*

1. Introduction

The Personalized Attention Records (PARs) of Linea Madrid are between the datasets made available by the local government of Madrid City in its open data portal https://datos.madrid.es/portal/site/egob. These records may be considered as a sort of counterpart from the call receiver's perspective of the Call Detail Records produced by telecom providers. The source for the PARs is the Customer Relationship Management (Buttle and Maklan, 2015), a front-end tool offering an interest oriented management solution. It gathers the data from different communications channels: the 26 citizen attention offices distributed by borough, the 010 phone number, the website chat, the Facebook account and the Twitter account @lineamadrid. The volume of the downloaded information, more than 44 million of records from 2014, cannot be processed using conventional statistical software and requires procedures specially developed for this purpose. Apache Spark (Zaharia et al., 2016), an open source analytics engine for Big Data processing has been used for the first steps of collecting and pre-processing data. Besides this, Python software (Van Rossum & Drake, 2009) and Scikit-learn (Pedregosa et al., 2011), a free software machine learning library for the Python programming language have been used for further calculations and analysis.

Each record contains a number of variables that have been changing through time. Some of these variables, such as the responsible worker or whether the issue has been addressed to another instance, are mostly interesting for administrative purposes. But there are other variables whose analysis may help to improve customer attention services. Besides the reception and register date, this paper focuses on the study of variables remaining through time, such as the topics that concern the citizens and the different channels dealing with the services.

Table 1. Examples of topic description variables in Personalized Attention Records.

Tipo 1	**Tipo 2**	**Tipo 3**	**Tipo 4**
Información general	Administración Pública	Administración estatal	
Movilidad	Madrid Central	Alta personas	
Identificación electrónica	Acceso a Carpeta Ciudadano	Alta	
Tasas e impuestos	IBI	Consulta/Información	Voluntaria
Cita Previa	Cita Previa	Asignar cita previa	
Movilidad	Multas	Pago con tarjeta	Voluntario
Padrón municipal	Justificantes empadronamiento	Volante empadronamiento	
Registro	Registro	Anotación	
Avisos	Avisos	Alta/Reiteración	

There is a specific variable reflecting the channel while the topic is spread out over four variables (tipo1 to tipo4). The target of using four variables to collect the topic seems to be obtaining a hierarchical description allowing for detailed information. But in practice the data have been filled in in various ways as can be seen in Table 1.

Daily indicators of the number of requests or questions received by channel and topic will be computed to provide an idea of the manner in which citizens' attention is managed by municipality services. For this purpose, the topic description variables need to be previously treated by Natural Language Processing tools.

The remainder of this paper is organized as follows. Next section describes the first steps of processing the records, making them homogeneous and obtaining a realistic classification of topics; section 3 presents and analyses the results obtained for the period between January 2014 and March 2020; and finally, a number of remarks and conclusions are presented in Section 4.

2. Processing of the Personalized Attention Records

The datasets including the records registered in the month are made available in the Madrid City open data portal after the end of each month. These files include the information of a number of variables varying in time making around 700 000 data points for each year and each variable.

The first step consists always on detecting and correcting possible logical inconsistencies in the data. For instance, for each new dataset, changes on date formats and variables appearing or disappearing are frequently found and must be previously detected to allow for subsequent homogeneous treatments. Likewise, once within the file, common spelling errors in string variables that have a fixed number of categories can be detected and corrected.

After the previous corrections, the selected data reflect the reception date, the channel receiving the request and the topic description variables. These last four variables (tipo1 to tipo4) must then be treated to obtain practical categories for classifying the topics which the requests refer to. For this purpose, some steps of Natural Language Processing (Jurafsky & Martin, 2008) have been carried out:

- Concatenation of variables tipo1 to tipo4 into one topic string, and conversion into lower case letters.
- Tokenization of the topic string into words.
- Elimination of duplicate words.
- Elimination of Spanish stop words (most common words that are filtered out).

These steps result in a list of significant words reflecting the topic for each record. A wordcloud is shown in Figure 1, which consists of a visual representation of the words in the topic descriptions where the size of each word is proportional to its frequency (Halvey and Keane, 2007).

Figure 1. Wordcloud for the words reflecting the topic for the Personalized Attention Records.

The first attempt to obtain a classification for the topics was done using Latent Dirichlet Allocation (Blei et al., 2003), a method for unsupervised classification that try to identify the topics best describing a set of documents. It builds a probabilistic model where each topic is characterized by a distribution over words, and documents are assigned probabilities to come from each topic. In the case of the PARs, the list of words for each record is too short, producing not interesting topics.

The next step has been finding the most frequent lists of one or more words by means of the Frequent Pattern Growth Algorithm, an efficient method of finding association rules (Witten et al., 2016). A classification on 15 classes has been established using these lists: *Bicycle*, *Culture and Sport*, *Elderlies*, *Electronic Signing*, *Environment*, *Employment*, *Foreigners*, *Health*, *Central Madrid*[1], *Mobility*, *Other*, *Penalties*, *Register*, *Taxes*, and *Urban Planning*. In this way, each record is classified in one of the categories depending on the words in its topic description.

[1] New big traffic-restricted area

3. Analysis of the results

For each record the data include now the date, the channel receiving the request and the topic it refers to. A first question to raise is about the relationship between channel and topic, whether there is any kind of association between these two nominal variables. The most popular measure of relationship between this type of variables is Cramer's V (Cramér, 1946) that takes values between 0 and 1, with values closer to 1 indicating a greater association. In this case its value is 0.20, reflecting a low-medium level of relationship.

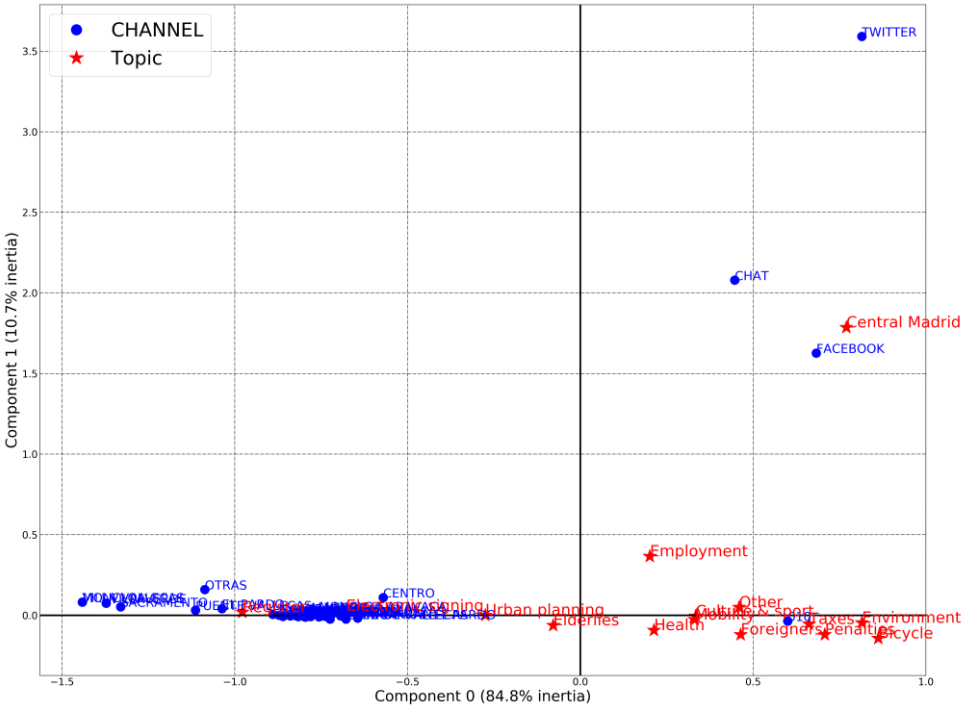

Figure 2. Principal coordinates for Channel and Topic categories.

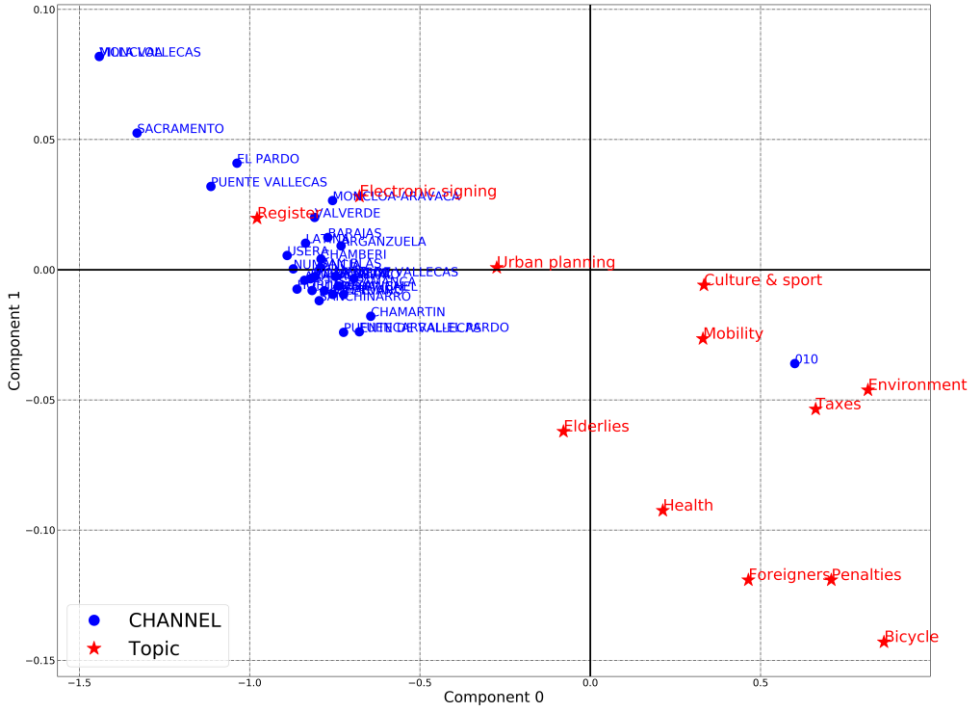

Figure 3. Principal coordinates for Channel and Topic selected categories.

Another way to study the association between channel and topic is the simple correspondence analysis (Greenacre, 2007), a graphical visualization of the rows and columns of a two-way contingency table as points in a low-dimensional space, where the positions of the row and column points are consistent with their associations in the table. The symmetrical plot including the principal coordinates for channels (rows) and topics (columns) appears in Figure 2. What can be said at first glance is that *Twitter*, *Chat* and *Facebook* channels have a similar profile of requests per topic, and that the requests with *Central Madrid* topic are more frequently received by these channels. Figure 3 amplifies Figure 2 to show other categories in more detail. It can be seen that the 26 borough offices have similar requests profiles and receive more frequently the requests related to *Register*, *Electronic signing*, and *Urban planning* while the *010* phone number receives mostly requests related to *Environment*, *Taxes*, and *Mobility*. The requests are unevenly distributed by channel, with the *010* phone number receiving more than 55%. The distribution per topic through all periods is shown in Figure 4. It must be taken into account that some topics such as *Central Madrid* and *Bicycle* have more recently appeared.

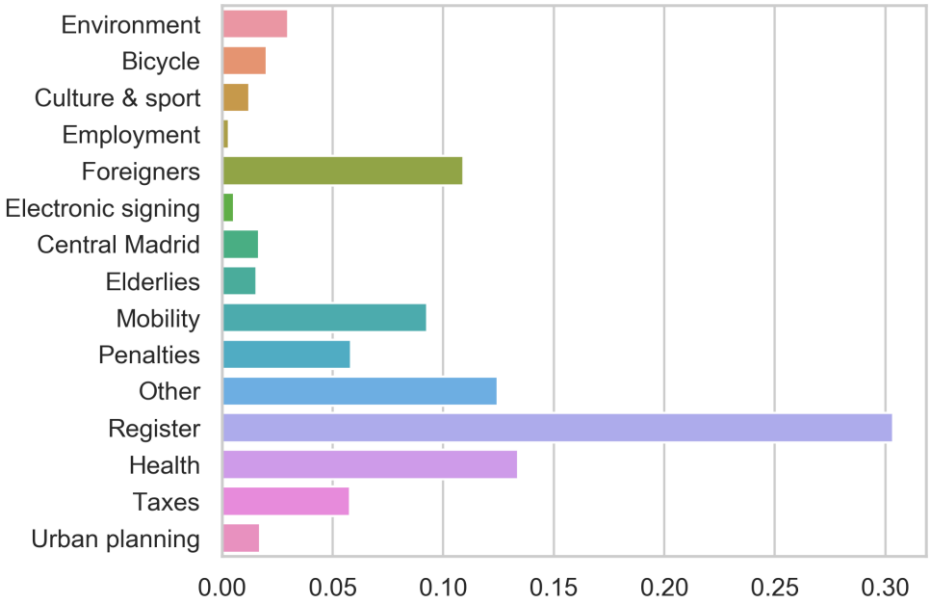

Figure 4. Proportion of requests per topic through January 2014 - March 2020.

The next step is to compute the proposed indicators. Figure 5 shows the time series of the total number of requests dealt with per day with a trend line, for the period between January 2014 and March 2020. Daily indicators for channels and topics have been similarly computed resulting in diverse rhythms, seasonal behaviors, trends, changes, and evolving behavior. The paper just presents some characteristics of the total series.

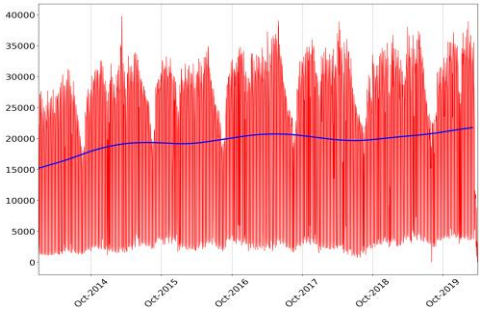

Figure 5. Total number of requests per day and trend line.

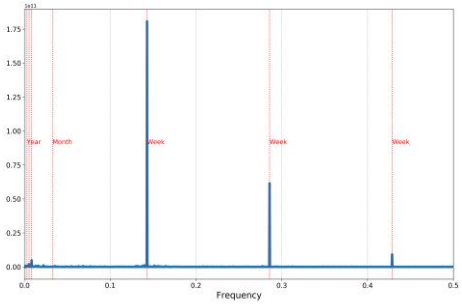

Figure 6. Periodogram spectrum estimates.

It can be seen in Figure 5 that the day-to-day movement has a lot of noise. There is also a strong pattern of seasonal decreasing in August and less markedly in December. Its periodogram spectrum estimated using Welch's method (Welch, 1967) is shown in Figure 6,

where the peaks in the spectrum indicate the frequencies of cyclical movements. The highest frequencies correspond to weekly periods, there are small frequencies for annual periods, and the frequencies are only just different from zero for monthly periods. Therefore, the most important cyclical oscillations correspond to weekly periods although these oscillations can hardly be seen in Figure 5 due to the big number of data.

To perform the seasonal adjustment of the series, a plugin of JDemetra+ version 3.0 (Eurostat, 2017) for the seasonal adjustment of daily and weekly series developed for testing purposes has been used. Boxplots of the weekly seasonality distribution are shown in Figure 7, where it can be seen that this seasonality is very stable and regular with decreasing growth the weekdays and decay on weekend. The annual seasonality average is shown in Figure 8. It results very irregular on its side because it has been calculated only with less than seven years and must be taken with caution. It reflects decays on periods matching dates of celebrations or holidays and days around them (first of May, May the 15, October the 12, first of November, …) and during the summer holidays (July - August).

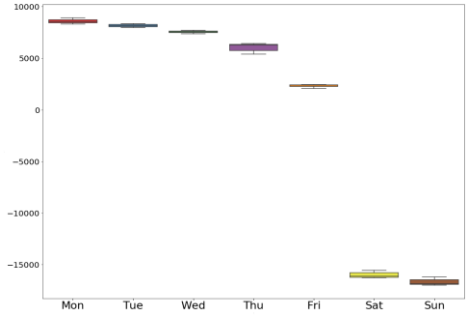

Figure 7. Weekly seasonality boxplots.

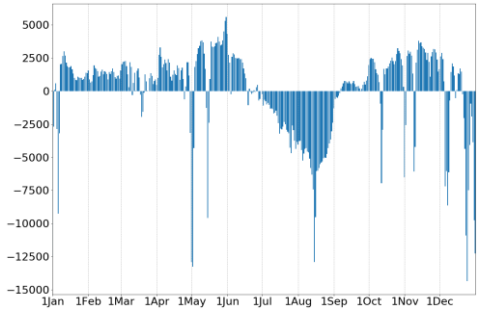

Figure 8. Annual seasonality average.

4. Last remarks

The previous sections have shown a procedure to build a classification for the topics to which the PARs refers to, and also how to compute daily indicators for the different channels and topics. The indicators can be analysed from the time series perspective to learn about the levels of workload in the channels, the seasonal behavior per topic and other issues.

There are many different ways in which this information may help to improve the attention to citizens. Although there have been shown just some examples for the global requests due to the limitation of space, other possible cases are: computing an approximate idea of the average level of occupancy of each channel during the week by calculating the average of the number of calls per day of the week and hour, and later dividing these averages by the maximum found number at this channel in an hour; executing a more precise natural language analysis of the typical issues requested by users within a topic to build a software application

for automating the customer service by conducting on-line chat conversations instead of providing direct contact with a live human agent.

References

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent Dirichlet Allocation. *Journal of Machine Learning Research*, 3 (4–5), 993–1022.

Buttle, F., & Maklan, S. (2015). Customer Relationship Management: Concepts and Technologies. NY: Routledge, Business & Economics.

Cramér, H. (1946). Mathematical Methods of Statistics. Princeton: Princeton University Press.

Eurostat (2017). JDemetra+ Reference Manual version 2.2. Retrieved from https://ec.europa.eu/eurostat/cros.

Greenacre, M. (2007). Correspondence Analysis in Practice (2nd ed.). London: Chapman & Hall/CRC.

Halvey, M., & Keane, M. T. (2007). An Assessment of Tag Presentation Techniques. Poster presentation at World Wide Web Conference 2007. Calgary, Canada.

Jurafsky, D. & Martin, J. H. (2008). Speech and Language Processing (2nd ed.). NJ: Pearson Prentice Hall.

Pedregosa, F., Varoquaux, G., Gramfort, A. , Michel, V., Thirion, B., Grisel, O., Blondel, M.,Prettenhofer, P., Weiss, R. , Dubourg, V. Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., & Duchesnay, E. (2011). Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research*, 12, 2825—2830.

Van Rossum, G., & Drake, F. L. (2009). Python 3 Reference Manual. Scotts Valley, CA: CreateSpace.

Welch, P. D. (1967). The use of Fast Fourier Transform for the estimation of power spectra: A method based on time averaging over short, modified periodograms. *IEEE Transactions on Audio and Electroacoustics* 15(2), 70–73.

Witten, I. H., Frank, E., Hall, M. A., & Pal, C. J. (2016). Data Mining: Practical Machine Learning Tools and Techniques (3rd ed.). Burlington, MA: Morgan Kaufmann.

Zaharia, M., Reynold, S. X., Wendell, P. Das, T., Armbrust, M., Dave, A.,Meng, X., Rosen, J.,Vernkataramen, S., Franklin, M. J., Ghodsi, A., Gonzalez, J., Shenker, S. & Stoica, I. (2016). Apache Spark: A Unified Engine for Big Data Processing. *Communications of the ACM*, 56(11), 56-65.