

LabUrbano Criminality

Ariza B. Francisco, Manchola M. Jofre, Ruiz S. Jairo, Clavijo Andrés, Guerra G. John

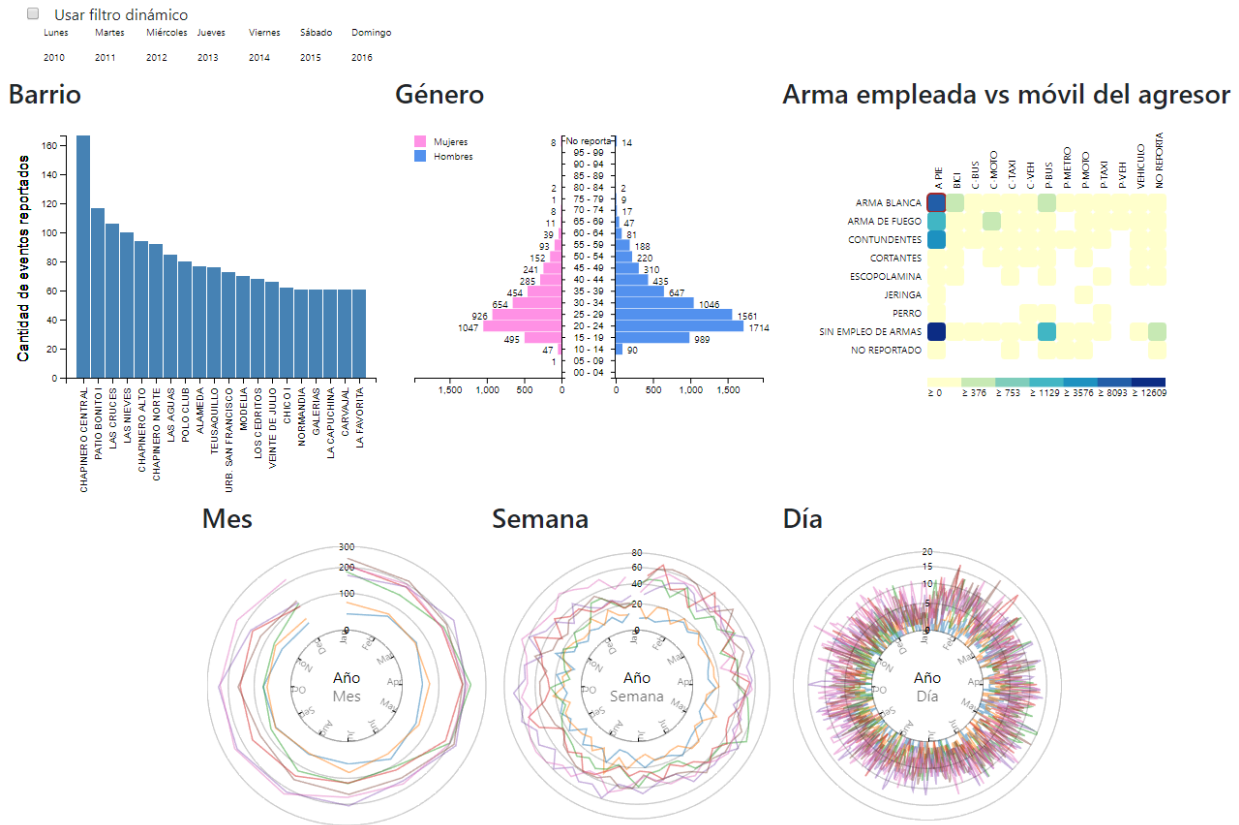


Fig. 1. Dashboard with interesting variables to discover insights from data on mobile theft.

Abstract—Mobile phone theft is a crime of high impact in the Bogotá city, the increase in the number of reported thefts, concerns the authorities and the people in general, however, the majority of the people is unaware of the areas in which this crime is presented, ignoring variables of time, mode and frequency of events due to the limited knowledge of the information and the little disclosure by the authorities. With official data of cell thefts reported between the years 2010 and 2016, it is proposed to implement a visualization that allows presenting and discovering insights for public use and ownership.

Index Terms—SIEDCO, cell phone theft, visualization, viz

1 INTRODUCTION

According to the last survey of Perception and Victimization in Bogotá carried out in September 2017 [2], 77% of respondents said they had been victim of the crime of *theft to person* and 40% of them *stole his cell phone*.

- Ariza B. Francisco student of Universidad de los Andes. E-mail: ja.arizab@uniandes.edu.co.
- Ruiz S. Jairo student of Universidad de los Andes. E-mail: ja.ruiz907@uniandes.edu.co.
- Manchola M. Jofre student of Universidad de los Andes. E-mail: ja.mancholam@uniandes.edu.co.
- Clavijo Andr Coordinator and Lead Researcher for Colombia — Data-Pop Alliance. E-mail: aclavijo@datapopalliance.org.
- Guerra G. John professor of Universidad de los Andes- E-mail: ja.guerrag@uniandes.edu.co.

Taking as starting point these alarming numbers, which worry the citizens of the community, authorities, victims and experts in the subject, the following questions arose: What is the neighborhood of Bogotá with the highest rate of cell phone theft? What could be the possible determining causes? In what areas of the city do you steal cell phones more frequently? What are the probable causes of cell phone theft? Is mobile theft higher in high-end devices, or do they steal without distinction? These are specific questions that require answers supported by data provided by an official entity such as the General Directorate of the National Police.

Based on these questions and with the participation of *Data-Pop Alliance*¹ as an expert user and main client, the state of the art of the

¹Data-Pop Alliance is a global coalition on Big Data and development created by the Harvard Humanitarian Initiative, MIT Media Lab, and Overseas Development Institute that brings together researchers, experts, practitioners, and activists to promote a people-centered Big Data revolution through collaborative research, capacity building, and community engagement. As of February

proposed problem was reviewed and an objective was defined to answer these and other questions that will be discovered with the analysis of the data and explorations with expert users. To achieve this goal, we propose in this paper as a work methodology, perform the cleaning and understanding of the data, define the users and tasks, define and develop an interactive visualization with filtering options that allows from a general question, to discover insights specific, combining two or more dimensions of the data.

2 STATE OF THE ART

The theft of cell phones is one of the most recurrent criminal phenomena in the Bogotá city of which has spread throughout the national territory, the alarming growth in the figures of this crime has served to consider it as an object of social study and economic, in order to be able to identify possible repercussions as a result of this illicit activity. Below, some visualizations of articles and investigations are analyzed:

- 1. Extinction of domain as a tool against cell phone theft in the Bogotá city. [4]

Task 1: Present the impact of the implementation of the extinction of the domain on the historical figures of mobile theft.

Modism Used: Bar graph (vertical and horizontal), lines.

Analysis: the visualizations support the proposed task, however, other types of graphics can be included to perform more comparison tasks (which is the main objective). See figures 2 y 3.



Fig. 2. Cell phones stolen by locations in Bogotá, 2011-2015.

- 2. Socio-economic, financial and legal impact of the licit and illicit cellphone commerce in Florencia Caquetá. [1]

2016, Flowminder Foundation has joined Data-Pop Alliance as its fourth Core Member.”

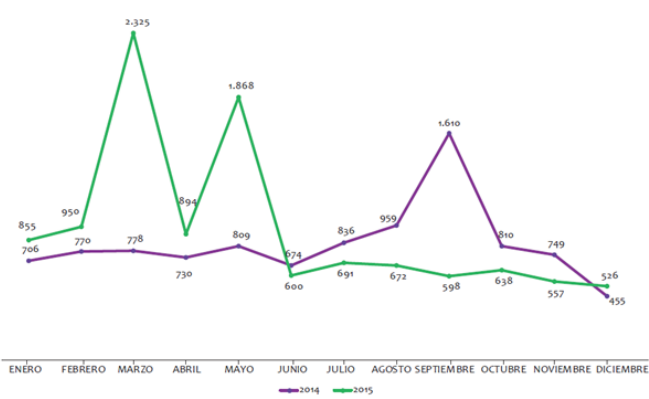


Fig. 3. Monthly behavior of cell phone theft in Bogotá, 2014-2015.

Task 1: Present the results of the survey conducted to the population of Florencia.

Modisms used: Bar graph (vertical and horizontal) and pie graph.

Analysis: The bar graphs are not ordered and in some cases the values presented are: eg, gender account and age account. See figures 4 y 5.

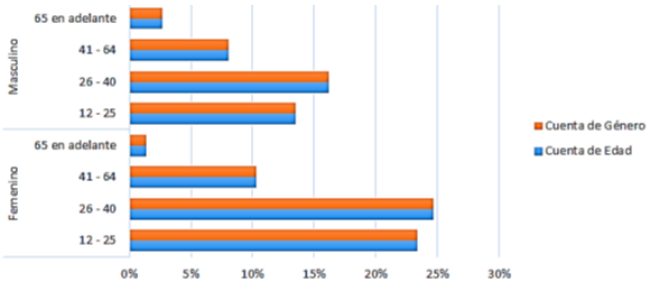


Fig. 4. Age of the victims of theft.

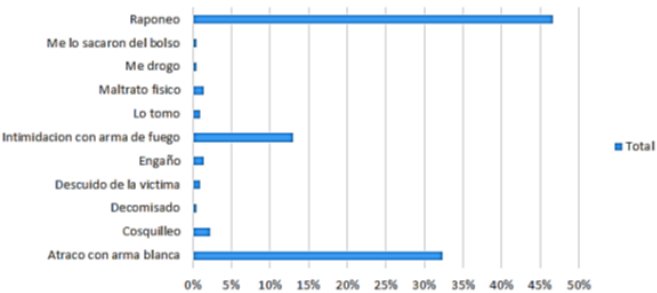


Fig. 5. Activity of the victim during the theft.

- 3. Heat map, cell phone theft in Bogotá. [5]

Task 1: Identify the localities and critical areas with the greatest number of thefts.

Task 2: Present the cell phone theft figures by location Modism used: Interactive map.

Analysis: it is a visualization oriented to the final user in which it is expected to acquire insurance for your cell phone, therefore the data is segmented by location. To make comparisons and a better visualization of data it is better to use bar charts (see figure 6).

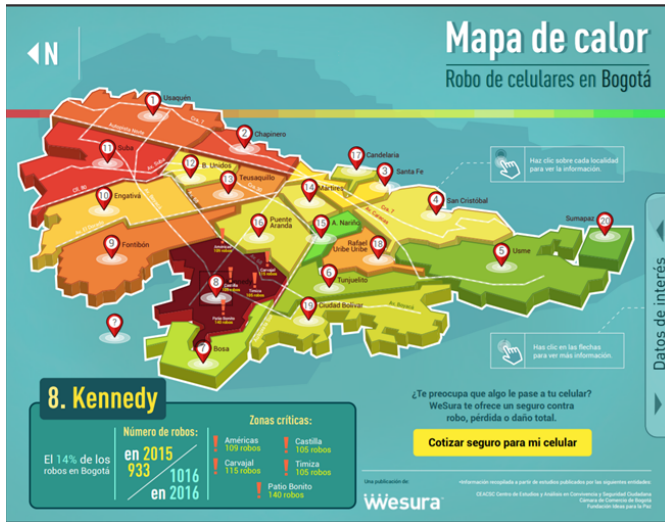


Fig. 6. Mapa de calor, robo de celulares en Bogotá.

3 CHARACTERIZATION

The characterization process was carried out based on the *framework* of Tamara Munzner [3] which is composed of the What, Why and the How. The What (abstraction of data) and the Why (tasks abstraction) are described below.

3.1 What

Dataset: Crime of theft to cell phone in the urban area of Bogotá. (see table 1).

Data source: has provided the dataset that it has consolidated on mobile theft in Bogotá, whose source, as information producer, is the police. Through the *Fundación Ideas para la Paz* (FIP), it was possible to obtain a dataset, which in addition to the attributes available in the open data portal of Colombia (<https://datos.gov.co/>), has the geographic information (latitude and longitude) of the event "theft to cell". This makes it possible to enrich the dataset with other geographical information such as neighborhood, police quadrants or police stations. It is worth noting that the geographic information is confidential, so the processed dataset, analyzes and visualizations will not show this information specifically.

Table 1. Relationship of Dataset Attributes and Types

Atributo	Tipo
Arma empleada	Categorico
Barrio	Categorico
Cantidad	Ordenado, Cuantitativo
Clase sitio	Categorico
Código DANE	Categorico
Departamento	Categorico
Día	Ordenada, cíclica
Empleado	Categorico
Escolaridad	Categorico
Fecha	Ordenado, Ordinal, Secuencial
Género	Categorico
Hora	Ordenado, Ordinal, cíclico
Línea	Categorico
Marca	Categorico
Móvil Agresor	Categorico
Móvil Víctima	Categorico
Municipio	Categorico
País Persona	Categorico
Timestamp	Ordenado, Ordinal, Secuencial
Zona	Categorico

3.1.1 Data cleaning

Taking into account that it is desired to perform an analysis of the data using the date of the theft as the main variable, 5,003 records were eliminated, given that they do not have information in the date attribute. Given that data from the city of Bogotá are used only, the variables Código DANE, Departamento y Municipio, are no longer relevant, since they become constant, therefore they will not be taken into account.

As a first analysis of the data, a profiling of the data was performed (see annex, report file in HTML format), where errors were found by completeness of the data, since they used special characters "-", or phrases such as NO REGISTRA, NO REGISTRARON, among others. Therefore, the data was homogenized and NO REPORTADO was used as a phrase to indicate the lack of information. In the same way, errors were found in the attributes of time, line and age, since they used data with different types or formats. Since the original dataset has an hour attribute and another date attribute, the data was derived generating a new variable called "TIMESTAMP", which records the date and time of the theft in a single attribute, which will facilitate the generation of visualizations.

3.1.2 Expert knowledge

Through *data-pop-alliance* we connect with the *Fundación Ideas para la Paz* (FIP) who have a dataset obtained from the police information system SIEDCO (Statistical, Delinquent, Contraventional and Operational Information System of the National Police), this dataset, in addition to the attributes presented above, has attributes of position, latitude and longitude, allowing to improve the quality of the data in terms of neighborhood or other jurisdiction of interest, such as localities, planning units or police quadrants, making a geographical intersection between the layer of crimes (with geometry point) and the layer of interest (with geometry polygon).

According to the FIP, the data registered in SIEDCO had a significant change in the way of recording the information in February 2014. Before this date, the police registered in SIEDCO the address in a text field and subsequently obtained its latitude and longitude by means of a geocoder, which generated a low reliability of the obtained position and many records could not be geocoded; after this date, the police had a map, in which it is possible to locate the point, eliminating the use of the geocoder and obtaining greater reliability in the geometry attribute of the dataset. For this reason, the FIP recommends using the series as of 2014 if geographical analysis is performed and taking the complete series (as of 2010) if alphanumeric analysis is performed.

3.2 Why

For the definition of why, the objective user of the project has been characterized together with *data-pop-alliance*, which has been defined with the following characteristics:

- Analytical citizen
- Interested in social problems.
- Example: User reading the "La silla vacía".
- A user able to understand a moderately complex visualization without becoming an expert.

The objective of the visualization is to allow the user to explore the data and **discover** characteristics of their interest.

3.2.1 In terms of Tamara

The main task *Analyze - Consume - Discover* and several associated secondary tasks are defined as:

- Discover** interesting features about mobile theft according to the place (neighborhood, site class), time and day of the week.

- Perform a time series analysis of the dataset and **Present** the distribution of cell phone theft as a function of temporary variables, whether it is time, day of the week, month, or even on an annual basis.
- Perform a geospatial analysis and **Identify** if there is a correlation between temporal and spatial variables, which allows identifying the place and time of the greatest occurrence of theft events.
- **Compare** cell phone theft trends between 2010 and 2017.

4 PROPOSED METHOD

In the process of defining the visualization, taking into account the tasks and data, the proposal of *HOW* was made (see figure 7) which consists of 5 idioms described below:

Modism 1

- Modism: map.
- Mark: Polygon (shape).
- Channel: Position (latitude, longitude) and color (saturation).
- Attribute: Number of thefts per neighborhood.
- Description: The number of cell thefts in the selected filter is represented with the saturation of the polygon filling.

Modism 2

- Modism: Bar Chart.
- Mark: Línea.
- Channel: Horizontal position, color (tone).
- Attribute: Amount of thefts to men and women by age ranges.
- Description: Represents the distribution of mobile theft by gender, discriminated by age range.

Modism 3

- Modism: Adjacency matrix.
- Mark: Rectangles (shape) and color (saturation).
- Channel: Horizontal position, vertical position, color (saturation).
- Attribute: Mobile of the aggressor and weapon used.
- Description: Represents the relationship between the mobile of the aggressor and the weapon used, with the saturation represents the number of thefts.

Modism 4

- Modism: Parallel coordinates.
- Mark: Lines
- Channel: Vertical position.
- Attribute: Time, day of the week, month and year.
- Description: Represents the relationship between the variables of the time in which the thefts were presented.

Modism 5

- Modism: Multi-line chart.
- Mark: Lines.
- Channel: Horizontal position, Vertical position, Color (tone)
- Attribute: Date, number of thefts, trends
- Description: Represents the behavior (feature) of the number of cellular thefts with respect to the time series (date)

5 RESULTS AND EVALUATION

Once the method was defined, it was proceeded to create what would be the first version of the mockup (figure 7), later it was subjected to user tests on which feedback was received from students of *Visual Analytics* and the project's client, The main results of the evaluation are listed below for each idiom.

Modism 1

- Users have difficulty finding specific neighborhoods, it is suggested to include textbox to search for locations by name.
- Use color saturation to express information (possibly number of theft events).
- It can be difficult to find the information of small neighborhoods, including zoom function on the map.

Modism 2

- Some users understand the divergence of data and the classification of information by age ranges, however, the different colors used confuse users. One color per gender should be used.
- Include some interaction / object to present the sum of events by gender (total men / total women) and general total.

Modism 3

- They consider interesting to present the variables of "mobile of thief" VS "used weapon", nevertheless a *heatmap* and not an adjacency matrix must be used.
- Use only one color to present the occurrence of events.
- You could consider adding *barcharts* to present the separated variables.

Modism 4

- There is confusion on the part of the users, given the complexity of the graphic. If implemented, the categorical variables must present the "values / categories" within the same column or in several columns in order to express the numerical values of the same.
- It is suggested to implement radial graphs to express the continuity (cyclical) of the temporary variables. <https://bl.ocks.org/mbostock/6fead6d1378d6df5ae77bb6a719afcb2>

Modism 5

- The users immediately relate the graphic to temporal data, however, when interacting, the purpose of the graphic is not clear.
- The interaction of filtering the data is not intuitive.
- The checkboxes are not visible enough.
- The graph works correctly to present the different trends, however, it does not facilitate tasks such as finding maximum values and their respective period, or comparison tasks.
- Validate if it is changed by another graphic.

General comments

- It is essential to include the titles of the axes in the graphs and their respective legends of the nomenclatures used.

Client comments

- It is important that you can easily answer the question: What is the probability of cell phone theft in a neighborhood given at a time XX and day of the week YY?

Dashboard Análisis Hurto Celulares - Ciudad de Bogotá

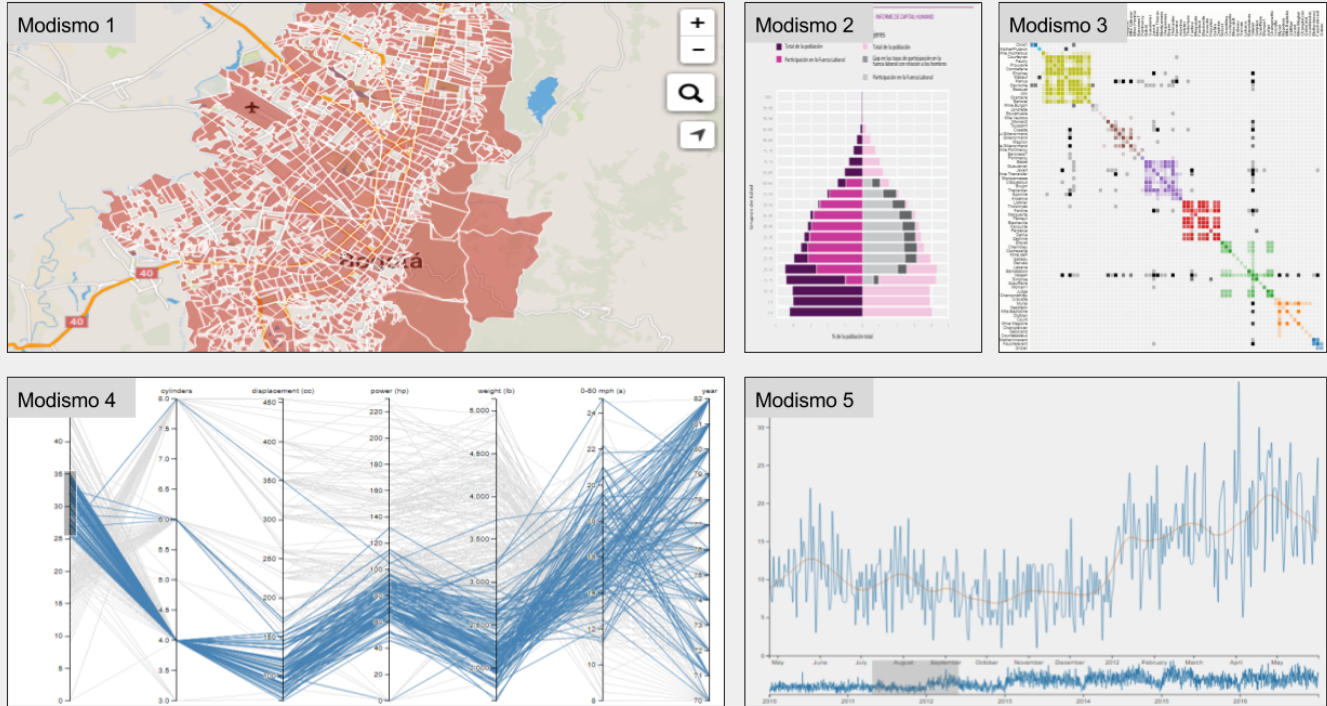


Fig. 7. Mockup version 1 of the proposed visualization.

Dashboard Análisis Hurto Celulares - Ciudad de Bogotá

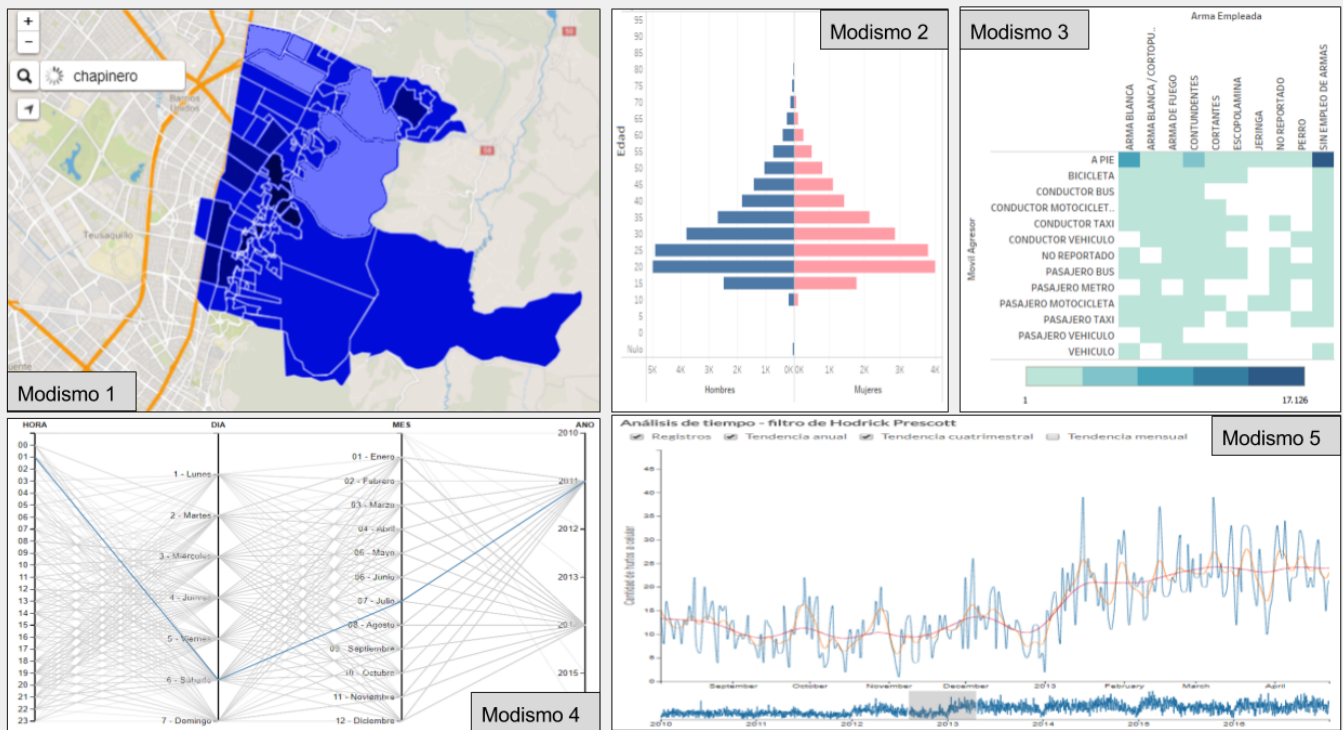


Fig. 8. Mockup version 2 of the proposed visualization.

- For the above I suggest that the dashboard have a box where is a clock or calendar.
- In order to better articulate the dashboard it is important to have a column with controls to select the main variables of crosses for the analysis. If you leave only the current squares the user can

lose a little.

- I like the map, the pyramid, the heat map and the analysis of time series. The graph of parallel bars is not so good, it is a good graph to see correlations, however, I think using it for only time variables is not recommended.

Finally, after evaluating the comments and making the adjustments to the visualization, the proposal was to develop a control board using *cross-filter*, the dimensions were defined in conjunction with data-pop-*alliance* during a review meeting of the descriptive analysis of the dataset.

As an approach and thinking about the attributes discussed, a second proposal was developed (see Fig. 8).

6 DISCUSSION AND CONCLUSION

The second version of the mockup was subjected to usability tests, as a result of the tests it was concluded that the idioms one, four and five had the following disadvantages:

- Modism 1: Finally we do not have the data of the x and y coordinates of each event, nor the description of the homologated neighborhood with the Bogotá neighborhood layer.
- Modism 4: The visualization requires a specific knowledge on the part of the user for its correct interpretation. It has occlusion problems due to the number of connections between parallel points.
- Modism 5: It is not clear to the user his interpretation with respect to the data filtered in the other idioms, mainly because trend lines are shown in a time range that make it impossible to combine variables such as year-month, day-hour.

Taking into account the above, the three mentioned idioms are modified by the following:

- **Modism 1**, allows to visualize the amount of thefts reported by neighborhood, it is filtered by one or more neighborhoods. (see Fig. 9)
- **Modism 4 y 5**, allow the user to view and select independently and easily the different magnitudes of time. (see Fig 10 y Fig 11)

6.1 Conclusions

The *dashboard* built with its 6 idioms allowed to discover several interesting insights among which the ones that are considered by the authors are significant:

- In 2010 more cell phones were stolen from men than women, however it is identified that year after year the figures of robbery to women have increased and for 2016 the number of robberies are similar between both genders.
- Men are more likely to be attacked with a knife in the event of cell phone theft.
- The majority of people who steal your cell phone using a firearm are men.
- The neighborhoods of the Chapinero area and the downtown area are the ones that present the greatest amount and frequency in mobile theft.

REFERENCES

[1] F. de Ciencias Económicas y Empresariales (Face). Universidad de Pamplona. Impacto socio-económico, financiero y legal del comercio de celulares lícito e ilícito en Florencia Caquetá. 16-1, 2016.

[2] C. de Comercio de Bogotá. Encuesta de percepción y victimización en bogotá. *The Visual Computer*, 1, Sept. 2017.

Barrio

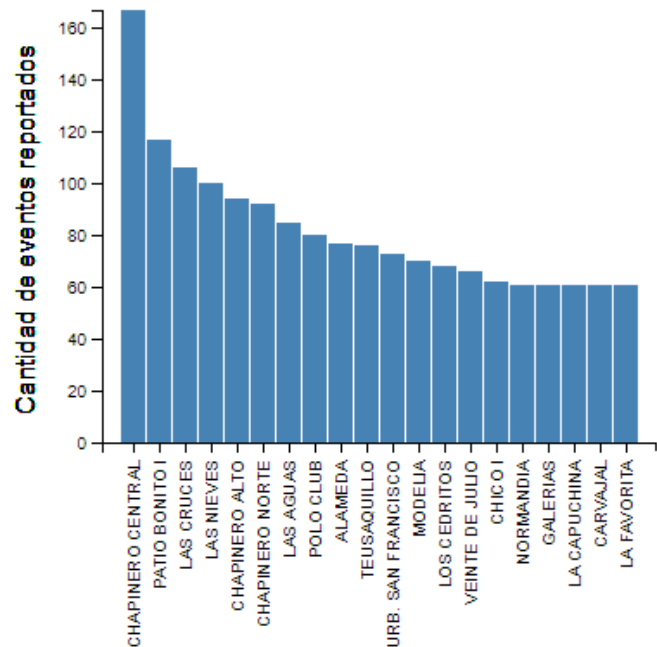


Fig. 9. Modismo 1 final version.

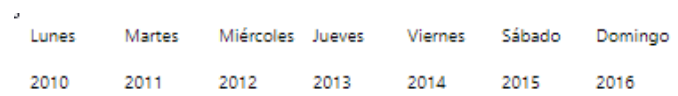


Fig. 10. Modism 4 y 5 final version.

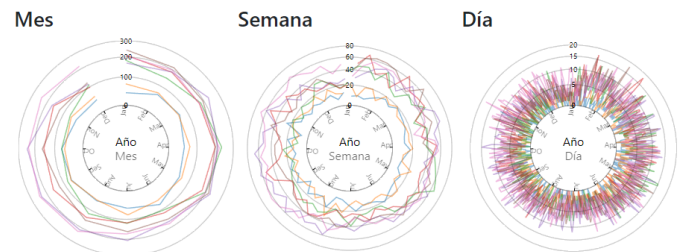


Fig. 11. Modism 4 y 5 final version.

[3] T. Munzner. *Visualization Analysis and Design*. Taylor and Francis Group, 6000 Broken Sound Parkway NW, SUite 300, 1nd ed., 2014. doi: ISBN 978-1-4665-0891-0

[4] L. Roa. Extinción de dominio como herramienta contra el hurto de celulares en la ciudad de Bogotá. *Revista Criminalidad*, 58-2, 2016.

[5] Wesura. Mapa de calor, robo de celulares en Bogotá. Wesura, Octubre 2017.