Smart Taxis

Heredia R. Fabián, Alvarado C. Manuel

Abstract— Mobility is a very important topic for Bogotá, over the years several administrations have centered in the task of improving it; another topic of interest that has emerged in the last years is the administration and regulation of taxi service and informal transportation service that has appeared in response to the issues with mobility, companies like Uber or Cabify for example, although they have not legalized their operation, they still offer a more ordered and safe service not only to passengers but to drivers also. District Mobility Secretariat has launched a project called Smart Taxis which consists of three main components: technological platforms (that includes mobile apps for drivers and users), Driver Unique Record and the modernization of the fleet. In this project it was used the data obtained from the technological platforms in order to analyze the behaviors and generate additional value.

Index Terms—taxis, Bogotá, mobility, transport, routes, ground transportation, transport analysis, visualization, data analysis.

Introduction

The District Mobility Secretariat (DMS) of the Bogotá Major City Hall is promoting the use of information technologies with the implementation of the Smart Taxis Project that seeks to "boost economy, transportation and technology sectors by activating innovation and production capacity of colombian companies and on the other hand to get the best individual public transportation service provision" [1]. This project implies the use of mobile applications by the taxi drivers, where this applications are developed by third parties and then validated by the DMS so information collected and generated by them is also recorded to the DMS' logs.

With the gradual implementation of this project, DMS is interested in evidencing everyday behavior of the taxis that belong to the project.

In order to achieve this the DMS hast provided a static set of anonymized data of services provided by the taxis linked to the project. This sets are distributed in several json format files.

Below it is presented the characterization and abstraction of the data following the Tamara Munzner Framework [2].

1 State of the Art

The DMS, taking into account the mobility of the city besides the passengers and drivers security and safety, assessed the mobility in other cities with a similar population, cities like Washington D.C., Mexico City, Mumbai, Querétaro and Minneapolis. From this cities, they checked their solutions of transportation by taxi, understanding their solution alternatives, they proceeded to assess colombian largest cities, taking into account collection prices and the number of habitants, they concluded the best way to formulate a collection rate was to do it in function of time and distance.

1.1 The DMS Current Solution

The DMS currently has as solution tool a visualization in Tableau, this solution pretends to show project indicators; its main task is to present tendencies in the taxi service of the city. There are other tasks which would be denominated as secondary and are to present the distribution

Fig. 1: Solución SDM, para el análisis de indicadores

of taxi services through time among others. It is kind of difficult to interpret information from some of the graphics due to the distribution, marks and channels do not allow to understand easily what is happening (Fig. 1).

2 What

There are two types of datasets, one contains information from some vehicles and its state, the states references whether the vehicle is Busy or Available for service. The second dataset describes the service requests and details whether the service was completed or interrupted before arriving the destination point.

2.1 Data Transformation

Original data received from the client corresponds to non valid json files, this is because they are lists of json objects that are not separated, therefore the first step was to repair the files so they could be processed and the information to use could be extracted. In the GitHub repository there are the python scripts used to perform such task under the **/py** directory.

2.2 DataSet 1

Dataset Availability: Static
Data Type: Items, Attributes
Dataset Type: Temporal

- *Heredia R. Fabián, Student at Universidad de los Andes. E-mail: fc.heredia10@uniandes.edu.co.*
- *Alvarado C. Manuel, Student at Universidad de los Andes. E-mail: ma.alvarado@uniandes.edu.co*
- *Nemocón F. Camilo, Project Client, District Mobility Secretariat. E-mail: cnemocon@movilidadbogota.gov.co.*
- *Guerra G. John, Professor at Universidad de los Andes. E-mail: ja.guerrag@uniandes.edu.co.*

Publish date Dec.4, 2018

Table 1: DataSet 1 Attribute Types

Attribute	Type	Description
tarjetaControl	Categorical	Taxi ID
fechaHora	Temporal	Date and hour of the measurement
latitud	Ordered Quantitative Diverging	Latitude of the geographical point the taxi was on the moment of the measurement
longitud	Ordered Quantitative Diverging	Longitude of the geographical point the taxi was on the moment of the measurement
estado	Categorical	State of the vehicle on the moment of the measurement. [D,O] that corresponds to Available and Busy
idCarrera	Categorical	ID of the service, only present during Busy state
h*	Ordered Ordinal	Hour range classification
origen*	Categorial	Binary field that determines whether the measurement is the point of origin of a service

* Derived fields.

2.3 DataSet 1.1

Dataset Availability: Dinamic (depends on filters)
Data Type: Items Attributes
Dataset Type: Table

Tabla 2: DataSet 1.1 Attribute Types

Attribute	Type	Description
estado	Categorical	State of the vehicle: Available or Busy
minutos*	Ordered Quantitativc Sequential	Number of minutes spent in a state

* Derived fields.

2.4 DataSet 2

Dataset Availability: Static.
Data Type: Table -> Items -> Attributes -> links
Data and Dataset Types: Network
Dataset Types: Networks.

Tabla 3: DataSet 2 Attribute Types

Atribute	type	Description
name *(node)	Categorical	Locality name
Node* (node)	Categorical	Node ID
Region *(node)	Categorical	Geographical dataset zone
Source* (link)	Categorical	Locality ID for the network origin
Target* (link)	Categorical	Locality ID for the network destination
Value* (link)	Ordered Quantitative Sequential	Number of services from an origin to a destination

* Derived fields.

3 WHY

There are identified the following tasks that should be executed with the purposed visualization:

3.1 Primary Task 1

3.1.1 Query, Identify - Trends

To evidence the moments when taxis belonging to the Smart Taxis Project are Available for service or Busy. For this it is understood that a taxi is Busy when transporting passengers from their origin to their destination point whether the service has been requested using a mobile app or taken in the streets and a taxi is Available for service when waiting for a service request.

3.2 Primary Task 2

3.2.1 Search, Browse - Outliers

To identify the city's localities where the most taxi services are originated and terminated.

3.3 Secondary Tasks

- To identify the amount of time (in minutes) a taxi spends in a determined state (available or busy). (Summarize -> Features)
- The geographical location of the measurements must be processed to identify the routes where a vehicle had a determined state. (Summarize -> Spatial Data).
- To identify the zones where the most taxi services are originated and terminated. (Query -> Compare -> Spatial Data)
- To compare the zones where the most taxi services are originated and terminated in Bogota. (Query -> Compare -> Spatial Data)
- To identify in a map the places where the services were originated and terminated. (Query -> Compare -> Spatial Data)
- The data from both sets must be preprocessed in order to achieve the primary and secondary tasks. (Analyze -> Produce -> Derive -> Features)

4 How

Following Tamara Munzner Framework [2], the how is defined as follows:

4.1 Primary Task 1

As the attributes to use are categorical (state), temporal, ordered, quantitative and cyclic (hour) and derived ordered, quantitative and sequential (quantity), the How is defined as:

- Encode:
  - Estado (state): Map – Color, hue.
  - Hora (hour): Arrange – Express.
  - Cantidad (quantity): Arrange – Express.
- Manipulate: Select.
- Facet: Juxtapose.
- Mark: Area.
- Channels:
  - Estado (state): Color, hue.
  - Hora (hour): Position.
  - Cantidad (quantity): Size.

4.2 Tarea Principal 2

As the data to use are mostly categorical as the transformation process the How is defined as:

- Encode: Separate, Order , Aling
- Manipulate: Select.
- Mark: Area.
- Channels:
  - Región (region): Node Color, hue.
  - Origin/Destiny: Node Horizontal Position.
  - Number of Services: Node Vertical Position.
  - Number of Services: Link Size.

4.3 Idiom Definition

As stated on the How definition, following expressiveness and effectiveness principles and according to user tests realized, discussed in numeral 5, the following idioms are defined:

- Primary Task 1: The idiom defined for this task is a bar diagram (histogram). In order to ease comprehension between the diagrams due to juxtapose use, it is implemented an interaction that through annotations displays the exact value for the same row in both diagrams.
- Primary Task 2: The idiom to use is a sankey diagram in order to easily visualize services distribution and relationships between nodes.

4.4 Associated Secondary Tasks

To summarize the amount of time taxis spend in each state, it is implemented another bar diagram (horizontal, in this case) so the data corresponds to each state displayed horizontally leveled and the codifications of the categorical (state) and derived, ordered, quantitative, sequential (minutes) attributes is defined as follows:

- Encode:
  - Estado (state): Map – Color, hue.
  - Minutos (minutes): Arrange – Express.
- Mark: Area.
- Channels:
  - Estado (state): Color, hue.
  - Minutos (minutes): Size.

To establish the routes where a taxi is in a determined state, it is implemented a map to georeference positions and its codification is defined as follows:

- Encode:
  - Estado (state): Map – Color, hue.
  - Latitude, Longitude: Arrange – Use.
- Mark: Shape (point).
- Channels:
  - Estado (state): Color, hue.
  - Latitud, Longitude: Position.

To identify the places where the most taxi services are originated and terminated, besides to be able to compare zones of this services:

- Encode:
  - Origin/Destiny: Map – Color, hue.
  - Latitude, Longitude: Arrange – Use.
- Mark: Shape (point or path)
- Channels:
  - Origin/Destiny: Color, hue.
  - Latitude, Longitude: Position.

To derive the data:

- In the case of the Activity section, once repaired the original data, it is executed the **preprocess-bars.py** script and its result is ordered by *tid* and *hour* columns (using excel due to the difficulty level of doing so in the same script because of the number of rows) and then to add the origin points it is executed the **add-origin.py** script.
- To the Use section, once original json files have been repaired, it is executed the **uso.py** script to generate the input file for the sankey diagram.
- In both sections it is generated a service in ArcGis for the map datasets.

5 Evaluation and Results

Once the initial idioms were defined with data characterization, it was generated a first version mockup and is tested with the Visual Analytics class mates; thanks to the feedback obtained it is generated a second version (Fig. 2, Fig. 3) and it is designed an Usability Test (Attachment 3) which is applied to 10 users (the project client included) and the summary of the problems found as follows:

5.1 Taxi Activity

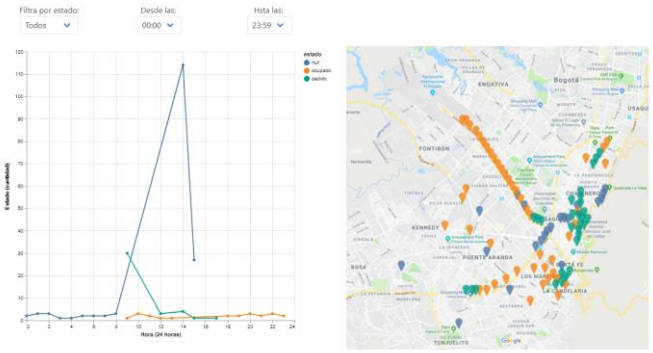

Fig. 2 – Taxi Activity Mockup

5.1.1 Idiom 1

- Users are confused about the number of taxis involved in the sample.
- Users find difficult to identify that the x axis corresponds to the hours of the day.

5.1.2 Idiom 2

- Users find difficult to relate the data between the idioms.
- Users would like to know the origin point of the services.

5.2 Uso de Taxis

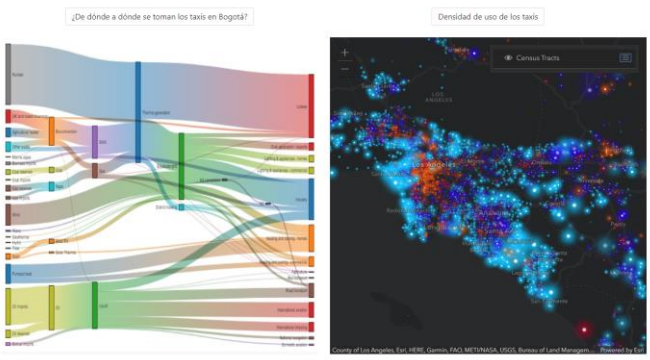

Fig. 3 – Taxi Use Mockup

5.2.1 Idiom 1

- Users are worried about what attribute is represented with the color.

5.2.2 Idiom 2

- Users find difficult to identify which points belong to origin and which to destiny.
- It is considered that a heat map could have a better performance.

Although the test results were satisfactory, it is obtained very relevant feedback from academy, users and project client and some adjustments are made to generate the final version of idioms (Fig. 4, Fig. 5, Fig. 6), for example, the first idiom of Taxi Activity is restated as a bar diagram (histogram) for each state.

6 Conclusions

Thanks to the feedback obtained during testing phase it is decided to change the idiom 1 from a line chart to a histogram by state and to implement a new idiom to represent the amount of time taxis spend by state, the last one was not contemplated in the initial versions of the mockups.

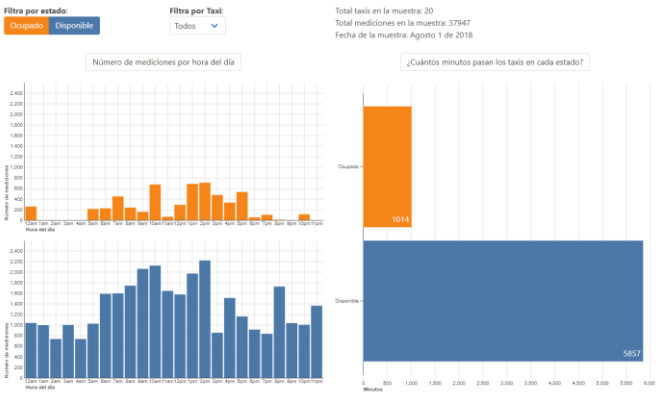

Fig. 4 – Final Idioms 1 and 2 (Taxi Activity)

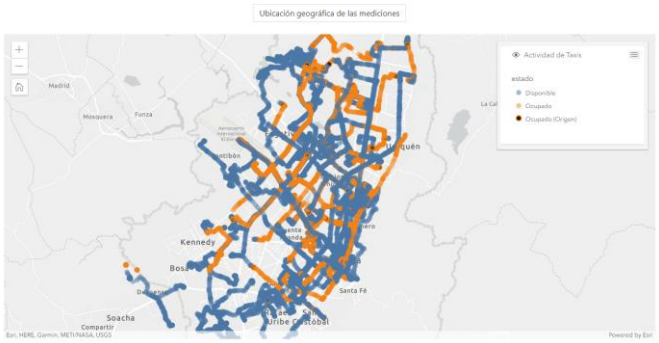

Fig. 5 – Final Idiom 3 (Taxi Activity)

With the idioms created for this section it is possible to state that:

- Due to the nature of the data one would expect the Busy state would be the dominant state, but Available state dominates by far even during peak traffic hours.
- Peaks in the graphics (where there are more active taxis and therefor more measurements) is between 10 and 11 am and between 1 and 3 pm, being specifically the highest between 2 and 3 pm.
- Despite there are available taxis there were no services in the extreme south of the city (localities like Ciudad Bolivar, San Cristobal and Rafael Uribe Uribe).
- It is very likely that some of the taxis may have lost connection or were moving while offline because in some cases there are not connected route tracks.

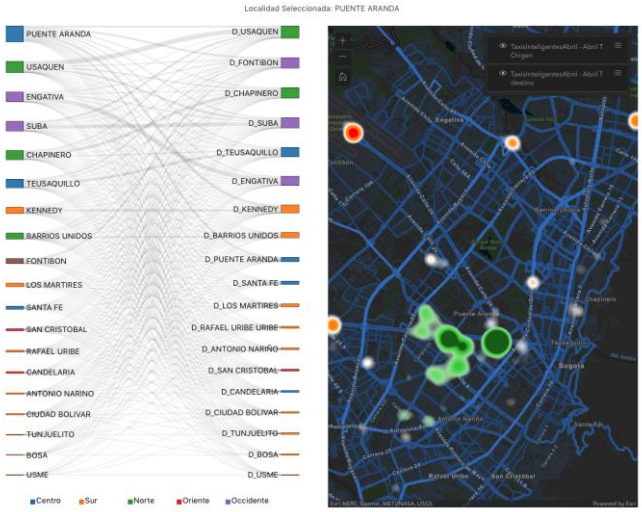

Fig. 6 – Final Idioms 1 and 2 (Taxi Use)

In the Taxi Use case, idioms allow to state that:

- During april and may 2018 most of the taxi services were originated from Puente Aranda and Suba respectively.
- During april and may 2018 most of the taxi services were terminated in Usaquén, Fontibón and Suba.
- Most of the taxi services are originated and terminated in the Center and North zones.
- With the map it is possible to note hot spots in where there could be defined "yellow zones" so the taxis are closer to their next service.

Final implementation is available in GitHub at https://fabianheredia.github.io/Taxis_Inteligentes/ powered by GitHub Pages. The source code can be found at https://github.com/fabianheredia/Taxis_Inteligentes.

References

[1] Secretaría Distrital de Movilidad. "Toda la innovación al servicio de los taxis inteligentes de Bogotá", disponible en http://www.movilidadbogota.gov.co/web/TODA%20LA%20INNOVACI%C3%93N%2C%20AL%20SERVICIO%20DE%20%20LOS%20TAXIS%20INTELIGENTES%20DE%20BOGOT%C3%81. 2018.

[2] T. Munzner. Visualization Analysis and Design. 2014.

[3]

Attachments

1. Mockup: mockup.pdf
2. Log and Schedule: bitácora-cronograma.xlsx
3. Usability Tests: Pruebas Proyecto Taxis Inteligentes.pdf