We know what you did - Congressman declared assets

Carlos J. Carvajal, Monica M. Carvajal and Marwan Mehenni

Abstract— Cuestion Publica is an independent journalist organization based in Colombia, focused on investigations concerning interactions between politics and economics, and more accurately in public topics such as health system management, public wealth, laboral laws, social conflicts and minorities. Their interest in such topics is guided to achieve public political control on governments and their members. Due to this interest, the "*We know what you did*" project was born, as an initiative to obtain political control on the representatives' wealth. Aiming, for instance, to identify trends, features outliers in the financial assets of the congress members [1], and furthermore to identify possible corruptions cases. Having these objectives in mind, we, as a team, consider Visual Analytics as the right approach to help Cuestion Publica fulfill their purpose, as we can provide a common language using the Tamara Munzner framework, and thanks to the enhancement that visualization produces in the human analysis capabilities.

Index Terms—Cuestion Publica, Colombia, political public control, congress, Visual Analytics, Tamara Munzner, Declaración de renta (DDR)

1 Introduction

In Colombia, the DIAN (Dirección de Impuestos y Aduanas Nacionales) is an official agency which controls and manages taxes on citizens and companies for incomes and goods, recollects tariffs for the flow of goods in and out of a country and manages taxes on betting games [2]. DIAN is under the direction of a bigger official entity called Ministerio de Hacienda which is itself under the control of the Presidency. The DIAN has standardized a process forcing individuals or companies with incomes coming from economic activities in Colombia, or those who possess assets (such as real estate properties or vehicles for instance). This categorization also includes citizens and companies from other countries. The process is known as **declaración de renta** (DDR) in which the aforementioned shall fill a form stating which goods they possess, how much they received as income and how much they have spent [3].

Public officers, as members of the government, are not exempt from this process, and in this case, the document is most of the time required by other entities, such as journalists, to exercise a political control. Cuestion Publica has required the DDR from twenty nine congress members in order to perform the aforementioned political control.

2 State of the Art

2.1 Similar works in other countries

Supervising Representative's earnings is nothing new, and in almost all democratic countries, their salaries and properties are public information. For example, such work has been done for the US houses of representatives, by the website Legistorm [1A], but only covers publicly accessible data. It doesn't cover private money movements, such as tax expenses or other various expenses.

Nonetheless, lobbying information is also available in the US, and it is possible to monitor

how much money the representatives have received from different lobbies. This information can be useful to better understand the stance of some representatives towards laws, and how they might be influenced. When the Senate voted the recent law about Net Neutrality, the website The Verge released an analysis about the telecommunications companies' contributions to senators [2A], and this helped citizens understand why the senators were so keen on passing this law that the public rejected massively.

2.2 What about Colombia ?

In Colombia, salaries are also publicly available, but what Cuestion Publica aims to address is to better understand what the Congressmen make with their money privately as a mean to public political control. This data is very complicated to obtain, because it relies on the will of the Congressman to turn in the documents, as well as a thorough manual examination of these documents. Cuestion Publica released a tool to easily visualize the documents provided by some congressmen [3A], but this doesn't include any Visual Analytics.

3 Users, Tasks and Data

As a first approach to the Tamara framework [4], we proceed to recognize three fundamental business entities, users, tasks and data. Below, we describe the users of the visualization, along with the goals and expectations. Then we describe the tasks the users would perform using the visualization, according to a previous meeting with them, and we propose some complementary tasks that will be discussed in future meetings. And finally, we describe the received data.

3.1 Users

We have identified two kinds of users: the client and the website users.

Cuestion Publica owns a website [5] where the team regularly updates the public on going investigations, as well as giving them access to documents recollected on such investigations. Cuestion Publica (CP) is run by three investigation journalists with some tech knowledge, as they recognize some terms as Javascript, programming languages and data cleaning. These three journalists are the clients, their expectation is to present a trend in the congressmen's wealth, looking to identify sudden rises in the incomes or properties.

The website users are our other kind of users. They are an heterogeneous group as we can't suppose any tech knowledge or even basic financial terminology knowledge. The website users, which themselves are 'clients' of CP, represent the final user, because even if the CP staff can understand a visualization containing financial, fiscal and tax terms, the page visitors are the ones who must be capable to understand the visualization and formulate their own insights.

As the final product must be displayed on the CP website (according to the staff requirements), our visualization must meet some non functional requirements such as low response times, and as mentioned above, the terms must be easy to understand for a non expert on fiscal jargon.

3.2 Tasks

To understand the tasks we first need to understand some terms present in the DDR.

- Patrimonio Bruto (PB): Represents all the assets possessed such as money in bank accounts, real estate, vehicles, stock shares, investments, etc.
- Pasivos (PV): The debts and obligations, as mortgages and loans or debts products of unpaid taxes.
- Patrimonio Liquido (PL): Represents the subtraction of PV from PB.
- Renta Bruta (RB): Incomes received in the year such as salary, viatic subsidies, fees, yields, and income from leases.
- Deducciones (DE): Some items that represent expenses and are subtracted from the RB to soften the paid tax, such as higher education payments, retirement payments and expenses for dependents such as underage children.

- Renta exenta (RE): Some items that represent incomes and are subtracted from the RB to soften the paid tax, such as a percentage of the earned yields.
- Renta Liquida Gravable (RLG): A earned quantity, base for the tax calculation, corresponding to RB = DE - RE
- Renta presuntiva (RP): The tax is calculated as we saw above according to the RLG, but may be the case of a low RLG but a high PL, this may be seen for individuals with many properties and money in the bank but no jobs, in this scenario the tax is calculated based on the previous years' PL.
- Impuesto a Cargo (IAC): This is the value that the congressman should pay.
- Total Rentas (TR): It's is the result of summarize Renta Líquida Gravable, Rentas Exentas, Total Ingreso No Constitutivo de Renta y Ganancias Ocasionales Gravables.

We will focus in just 3 variables PL, TR and IAC. The reason is that we know that if we understand this, we could response to the principal question of CP.

CP, as a primary task, want to visualize the trends on the PL. A rise in the PL as we saw previously, may indicate the acquisition of properties and the increase of money in the bank, or may indicate a decrease in the PV item.

We propose as complementary tasks:

1. To present trends of the average of TR, PL and IAC for all congress.
2. Compare and present the values of TR, PL and IAC of the difference (variance) and the total values by year and by congressman.
3. To present the distribution of the members of congress in terms of TR, PL and IAC.
4. Summarize the numbers which represents the results of the CP research.
5. Compare the different trends of the most importants variables which compose DDR.
6. Identify outliers in the DDR of each congressman or congresswoman.
7. Compare the members of congress by the most representative variables.
8. Lookup, locate and browse where each congressman is in each analyze variables.
9. To present the distribution of the most important variables of DDR.
10. Count the number of DDR by year that CP has to present the distribution of the information.

3.3 Data

CP asked for the DDR of twenty nine congress members, and from this quantity received documents for seventeen congress members. On these seventeen, CP have more than one DDR for 7 congress members, for the other ten, only one DDR is available. Having more than one DDR in consecutive periods of time is necessary to present trend analysis.

The data has been digitized in Excel files, with a first structure proposed by CP which may ease the analysis process. For the moment, only three congress members have all their DDRs in Excel format, the other ones are in progress.

Año		2014
Total gastos de Nomina	$	-
Aportes al Sisitema de Seguridad Social	$	-
Aportes al Sena,iCBF, Caja de Compensación	$	-
Total Patrimimonio Bruto	$	246,813,000
Deudas	$	224,149,000
Total Patrimonio Liquido	$	22,664,000
Salarios y demas pagos laborales	$	145,912,000
Honorarios, Comisiones y Servicios	$	-
Interes y Rendimiento Financiero	$	169,000
Otros Ingresos (Venta, Arreindo, Ect...)	$	-
Total Ingresos Recibidos por Concepto de Renta	$	146,081,000

Fig. 1. Excerpt of a DDR in Excel format, focus on PL

Gastos de Nomina Incluidos los aportes a seguridad social y Parafiscales	$	-
Gastos de Nomina Incluidos los aportes a seguridad social y Parafiscales	$	3,306,000
Deducciones inversión en activos fijos	$	-
Deducciones por pagos de intereses vivienda	$	1,131,000
Otros costos y deducciones	$	8,874,000
Total Costos y Deducciones	$	13,311,000
Renta líquida ordinaria del ejercicio	$	131,186,000
Renta Presuntiva	$	2,628,000

Fig. 2. Excerpt of a DDR in Excel format, focus on DE

4 Applying Tamara Framework

4.1 What

4.1.1 Data and Dataset types

The data is a table (which one has items and attributes, these will be describe later), in general, with temporal data. The temporal data will be the central point in this work. In addition, we will use some tables that don't represent temporal data, but instead contains data just for one period (as we have most of the DDR for 2016), this data will be useful to work on secondary tasks such as wealth static analysis. It'll be necessary to derive and obtain other datasets which would allow for a distribution analysis and a comparison of the different components of the DDR.

We derive a table with temporal data that includes the number of congressional seats by year and by chamber.

4.1.2 Attributes types

Table	Attributes	Type
DDR by Year (T1) (Temporal data)	DDR Variable	Categorical
	Year	Ordered, Ordinal, Quantitative, Sequential
	Congress member	Categorical
	Value	Ordered, Ordinal, Quantitative, Sequential
Number of congressional seats by year and by chamber (T2) (Temporal data)	Year	Ordered, Ordinal, Quantitative, Sequential
	Chamber	Categorical
	No. seats	Ordered, Ordinal, Quantitative, Sequential
DDR by Congress member (2016) (T3)	DDR Variable	Categorical
	Congress member	Categorical
	Value	Ordered, Ordinal, Quantitative, Sequential
Average of principal variable (i.e. PL, RL) by year (T4) Temporal data derived from T1, the objective is summarizing the data to analyze a global view.	Year	Ordered, Ordinal, Quantitative, Sequential
	Average PL (RL, PB, etc.)	Ordered, Ordinal, Quantitative, Sequential
	Number of congressman (this number is equal to the number of DDR that we have by year)	Ordered, Ordinal, Quantitative, Sequential
Results of CP research (T5)	Name of variable	Categorical
	Value	Ordered, Ordinal, Quantitative, Sequential

Table 1. Datasets, attributes and types

4.2 Why

Based on the tasks that we previously exposed, now we will describe those in a formal way used Tamara's Framework.

Primary tasks:

1. Compare (present) trends: Visualize the trends on the PL by congressman
2. Summarize distribution: Count the number of DDR by year that CP has to present the distribution of the information.
3. Identify and compare (present) outliers: Identify and compare (present) outliers on the relation between the earns of one year and the wealth of the next one.
4. Summarize (present) features: Summarize and present the numbers which represents the results of the CP research.

Complementary tasks:

1. Compare (present) trends: To present trends on average of PL, IAC and TR.
2. Compare and present features: Compare and present the values of the difference (variance) by year and by congressman for the most important variables like PL.
3. Identify outliers: In specific variables like PL, TR and other derived variables from the original data, of all congress members and in the DDR of each congressman or congresswoman.
4. Derive average, new variables like pocket change, variance and difference for the PL, TR and IAC for the congress members in same periods.

4.3 How

The solution is divided in 5 sections, each one representing a primary task. Each sections has an specific objective:

1. Section 1: Summarize and present the distribution of the information that CP received. (chart 1)
2. Section 2: Summarize, present and understand the trends of wealth our principal variable. (chart 2 to chart 4)
3. Section 3: Compare and understand the earnings and the taxes without discounts,

our second important variables. (chart 5 and chart 6)

4. Section 4: Compare and identify outliers based on derived data which contains new variables that represents difference and variations of the principal data. (charts 7 to chart 9)
5. Section 5: Summarize, present and understand the the effort and results of the CP's research. (chart 10)

Fig. 3. Final Solution. Section 1

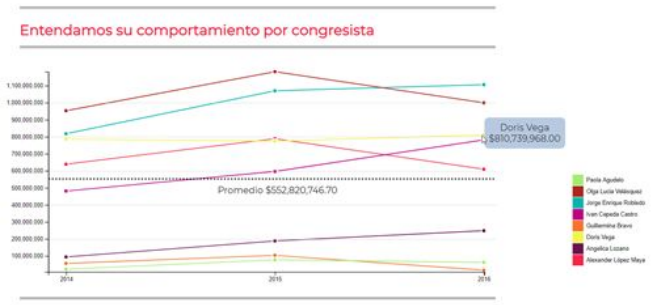

Fig. 4. Final Solution. part of Section 2

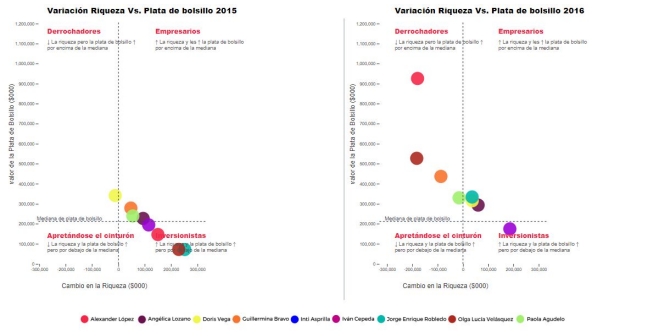

Fig. 5. Final Solution. part of Section 3

4.3.1 Information DDR

In this section our objective is response to the primary task 2 and the complementary tasks 1,4 and 7.

- Chart 1:
  - Encode: Express number of DDR (congressman) with horizontal and vertical spatial position and point marks. The horizontal position is determined by the year and express time.
  - Map: Use color hue to represent each congressman.
  - Manipulate: Navigate by point and year to read a tooltip that show the congressman represented by each point. Select a point to filter the information and focus in just one congressman.

4.3.2 Trends

- Chart 2 & 3: Contains a line chart by variable.
  - Encode: Line marks, express value attribute with aligned vertical position, express year which represent time across horizontal position and use the name of the variable as categorical attribute to identify each one by color - hue, the
  - Manipulate: Navigate by trends.
  - Map: Use color hue to represent each variable.

4.3.3 Variable Comparison

- Chart 4,5 & 6:
  - Encode: Line marks, express value attribute variation of variables with aligned horizontal position, separate by the name of the congressman (as a categorical attribute) as key attribute with vertical position. Use color - hue to differentiate between the period of the variance.
  - Manipulate: Navigate by each bar to read a tooltip that show the variance for each period and congressman.
  - Facet: Superimpose an average line.

4.3.4 Derived data

- Chart 7:
  - Encode: Line marks, express value attribute derived from the original data pocket change, separate by the name of the congressman (as a categorical attribute) as key attribute with horizontal position. Use color - hue to differentiate between the period of the pocket change.
  - Manipulate: Navigate by each bar to read a tooltip that show the exact value for each data point.
- Chart 8:
  - Encode: Use point as marks, express values of "Cambio en la riqueza" with horizontal position and express values of "Plata de bolsillo" with vertical position. Each point mark represent a congressman
    Line marks to express point 0 in horizontal position and the median in the vertical position.
  - Map: Use color hue to represent the categorical attribute congressman.
  - Manipulate: Navigate by each point to read a tooltip with addition information.
  - Facet: Same encoding, each scatterplot show subset of data with a specific period. Small multiples, use juxtapose views to support higher-precision comparison between both periods. Superimpose lines.
- Chart 9:
  - Encode: Line marks, express value attribute of each variable (Variación porcentual) with aligned horizontal position, separate by the name of the variable (as a categorical attribute) as key attribute with vertical position. Use color - hue to differentiate between the congressman selected and the total.
  - Manipulate: Navigate by each bar to read a tooltip that show the variance for each period and variable.
  - Reduce: Filter values by congressman.

4.3.5 Effort and next steps

- Chart 10:
  - Encode: Human shape as node with no links.
  - Map: Use luminance to represent 0 or 1.
  - Manipulate: Navigate use the buttons "back" and "next".
  - Reduce: Filter data to show each data subset use on it.

5 First Results - Analysis

5.1 The Data

As said, access to the data for a specific Congressman relies on their good will, meaning that they might not share all their data the same way. On average, Cuestion Publica only recovered 3 Declaracion de Renta per Congressman. This implies that the data isn't the same for each Congressman, especially temporarily : some share their data starting in the early 2000's, other only share theirs in the 2010's. This will be a challenge for representing the data : for some years, we won't even be able to compare two congressmen's data. We have to consider multiple options : only featuring periods where we have enough info to compare, or only featuring the congressmen for whom we have the most info.
The initial data we received was disorganized and needed a lot of formatting. We received 14 Excel sheets (one for each congressman) that looked like the Fig 6.

	A	B	C
1	Año	2006	2007
2	Total gastos de Nomina	$ -	$ -
3	Aportes al Sistema de Seguridad Social	$ -	$ -
4	Aportes al SGSS a cargo del empleado (75)	$ -	$ -
5	Aportes al Sena, ICBF, Caja de Compensación	$ -	$ -
6	Total Patrimonio Bruto	$145 649 000	143 450 000
7	Deudas	$7 101 000	4 513 000
8	Total Patrimonio Liquido	$138 548 000	138 937 000
9	Salarios y demas pagos laborales	$247 816 000	248 129 000
10	Ingresos recibidos como empleado (33)		
11	Honorarios, Comisiones y Servicios		
12	Interes y Rendimiento Financiero	$663 000	105 000
13	Otros Ingresos (Venta, Arriendo, Etc...)		
14	**Total Ingresos Recibidos por Concepto de Renta**	$248 479 000	248 234 000
15	Dividendo y Participativos		
16	Donaciones		
17	Pagos a terceros por alimentación		
18	Otros ingresos no constitutivos de renta		
19	Ingresos no consultivos de renta	$17 275 000	18 587 000
20	**Total Ingresos Netos**	$231 204 000	229 647 000
21	Deducciones inversión en activos fijos		
22	Deducción por dependientes económicos (48)		
23	Deducciones por pagos de intereses vivienda		

Alexandra López Maya | Angelica Lozano | Paola Agudelo | Ivan Name Vasquez

Fig. 6. Data format example

We needed to determine which cells contained the info we needed. This was provided by the client, along with a contextual summary table, which helped us understand what data we had for each congressman, and if the congressman is a representant from the Congress or the Senate.

In the process of cleaning and formatting the data, the tool *Alteryx* [4A] was used. It helped us to homogenize the data, and made it available for us to work with.

5.3 Analysis with the Client

We had in total 5 meetings with the client : 3 with the principal client and two with their financial analyst. Their principal concerns was about the use of the technical idioms and language and the importance to make something to be used by their users, so the necessity of have a good narrative, usability and to help to the user understand the whole topic.

5.4 Usability Test

We make two rounds, with different users, using improved mockups, and adding some storytelling. This usability test was performed on Sway[1B] and all the mockups were made using powerBi.

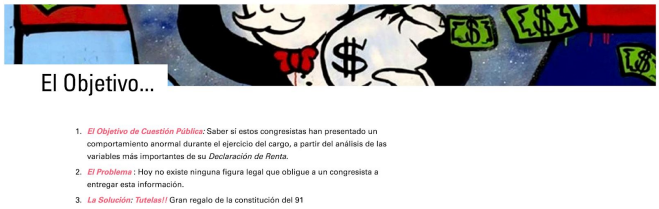

Fig. 7. An example of the added storytelling.

We test our mockups with 9 different users, using a scale from 1 (lowest) to 7 (highest). We ask our users to rank the idioms according to some tasks. For example, we ask the users to identify the years with the highest and lowest processed DDR. In general we achieve scores above 4 for all the idioms.

Usability Test Results		
Total number of responses		9
Average score on tributary knowledge		4.44/10
Gender distribution	Female	7
	Male	2
Average age		35,2
Education Level	University - College	6
	Postgraduate	3

Table 2. Demographic data on usability tests

5.5 Improvements after test

According to the users feedback [1B], we identified some changes on the idioms and in the whole site in general, we collect over 20 possible modifications, having as metric the frequency of coincidence of such suggestions, the input from the CP staff and our own opinion as a team, we prioritize the following improvements:

- Eliminating unnecessary juxtaposition.
- Remove some unwanted charts.

- Use a more informal language, avoid the technical words to name the variables.
- Increase font size.
- Use standard hue scales for categories.
- Use the photos of each congressman as a shape instead of the points.

This improvements were made in the last development stage.

6 Conclusion

Presenting the results of the CP research implies different challenges, one of the biggest is to use idioms that could be understood by any user of the website, regardless of their previous knowledge of the subject. Another challenge, since there is little information, was being able to compare the behavior between the congress members to identify outliers. We overcame this and our visualizations provide clear insights.

As a result from the usability tests, we have defined the changes on the actual mockups and the whole storytelling, this is an important input for the building of the final visualizations.

These visualizations provided us with interesting insights, including :

- Some congressmen "forgot" to submit their info for some particular years.
- Some congressmen such as A. Lopez Maya or O. Lucia Velasquez have some explaining to do regarding their "pocket change"...
- We now know that O. Lucia Velasquez and J. Robledo are the ones with the most global wealth.
- I. Cepeda was the one with the highest positive evolution in his wealth
- Even though the amount of work to get this data was important, we are still missing info for 96% of the congress...

7 References

[1] Congreso Visible, "¿Qué es el Congreso de la República?", https://www.congresovisible.org/democracia/congreso/, Consulted on 16/10/2018

[2] Direccíon de Impuestos y Aduanas Nacionales : https://www.dian.gov.co - Consulted on 16/10/2018

[3] Declaracíon de Renta form, DIAN, 2018, available at : https://www.dian.gov.co/atencionciudadano/formulariosinstructivos/Formularios/2018/Formulario_210_2018.pdf

[4] T. Munzner, *Visualization Analysis and Design*, CRC Press, 2014.

[5] Cuestion Publica : https://cuestionpublica.com/ - Consulted on 16/10/2018

Technologies used in this project :

- ❏ d3.js : https://d3js.org/ (version 5)
- ❏ Vega-Lite : https://vega.github.io/vega-lite/
- ❏ Tableau : https://www.tableau.com
- ❏ Power-bi : https://powerbi.microsoft.com/
- ❏ JavaScript: https://www.javascript.com/

References step 2 : [A]

[1A] Salary Rate Analysis, legistorm, https://www.legistorm.com/salary/analysis.html - consulted on 30/10/2018

[2A] Congress took $101 million in donations from the ISP industry — here's how much your lawmaker got, T.C Sottek, The Verge, https://www.theverge.com/2017/12/11/16746230/net-neutrality-fcc-isp-congress-campaign-contribution - consulted on 30/10/2018

[3A] Sabemos lo que hiciste la legislatura pasada ¡Recargado! , Cuestion Publica, https://cuestionpublica.com/sabemos-lo-que-hiciste-recargado/ - consulted on 30/10/2018

[4A] Alteryx, https://www.alteryx.com/

[5A] https://cjcarvajal.github.io/cuestion-publica/

References Usability test [B]

[1B] https://sway.office.com/HzxxjPfaCFdcvEEa?loc=swsp