A visual analytics framework case study: Understanding Colombia's National Administrative Department of Statistics datasets

Pierre Raimbaud[1*,2*][0000-0002-5584-8100], Jaime Camilo Espitia Castillo[1+][0000-0002-7261-3944], John A. Guerra-Gomez[3#][0000-0001-7943-0000]

[1] Systems and Computing Engineering, Imagine Group, Universidad de los Andes, Bogota, D.C., Colombia
*:p.raimbaud@uniandes.edu.co
+: camilospn@gmail.com
[2] LiSPEN, Arts et Métiers, Institut Image, Chalon-sur-Saone, France
*: pierre.raimbaud@ensam.eu
[3] Northeastern University, San Jose, California
#: john.guerra@gmail.com

Abstract. In a world filled with data, it is expected for a nation to take decisions informed by data. However, countries need to first collect and publish such data in a way meaningful for both citizens and policy makers. A good thematic classification could be instrumental in helping users to navigate and find the right resources on a rich data repository, such as the one collected by the DANE (*Departamento Administrativo Nacional de Estadística*, i.e. the Colombia's National Administrative Department of Statistics). The Visual Analytics Framework is a methodology for conducting visual analysis developed by T. Munzner et al.1 that could help with this task. This paper presents a case study applying such framework conducted to help the DANE to better visualize their data repository, and also to understand it better by using another classification extracted from its metadata. It describes the three main analysis tasks identified and the proposed solutions. Usability testing results during the process helped to correct the visualizations and make them adapted to decision-making. Finally, we explained the collection of insights generated from them.

Keywords: Visual Analytics, Data Repositories, Open Data.

1 Introduction

The DANE (Departamento Administrativo Nacional de Estadística, i.e. the Colombia's National Administrative Department of Statistics) is the Colombian public organization responsible for collecting, analyzing and distributing the country's national statistics. The total amount of data that this institution owns is one of the largest in the country (among public institutions), since it periodically gathers information about all the major topics, from population statistics, to public access to services, among many others. Because of this, one of DANE's main goals is that

public policies in Colombia become more data-driven [1]. However, this is rarely the case, as public institutions suffer data availability and/or classification issues. Aware of these issues, the DANE wants to improve its data management, by applying visual analytics methods, resulting in building and delivering better tools to the public policy-making structures.

Concretely, the DANE owns data coming from both administrative records (called administrative registers (AR) afterwards) and the derived statistical analyses conducted on them i.e. statistical operations (called statistical operations (SO) afterwards) and f. This paper addresses these two main types of data. Note that here we will not consider the final data collected by the DANE when they apply the questions or requests contained in an administrative register or a statistical operation, but we will consider the characteristics of the administrative registers and the statistical operations themselves, meaning their attributes and characteristics, in other terms all their metadata. So, our original data will be the DANE inventories of statistical operations and administrative registers. Fig.**1** illustrates this main distinction.

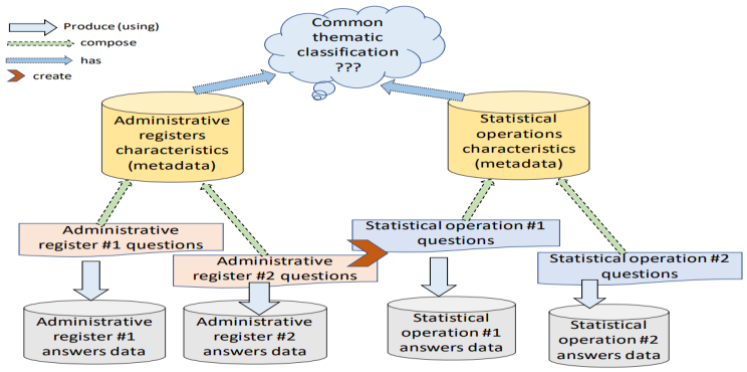


Fig. **1.** DANE's data distribution and explanation of the problematic: does a new thematic classification coming from the metadata of the administrative registers and statistic operation exist?

Based on these considerations, the main objective of this project is to build a tool to understand and visualize the DANE data, particularly by organizing it through the topics and keywords present among the metadata of different groups of statistical operations and administrative registers, ultimately allowing decision-makers to have right overviews of topics and to find which statistical operations and administrative registers are related to one in particular, thanks to this classification (see Fig.**Error! Not a valid bookmark self-reference.**).

2 Related work

2.1 Tamara Munzner's framework

For this project, we used Munzner's visualization framework [2] to abstract and understand the data, the users' tasks and to choose the best idioms that allow the users to complete these tasks. It has three dimensions: the WHAT, the WHY and the HOW.

WHAT: It refers to the available data for the visualization. The basic abstractions of the dataset arrangements are tables, networks, fields and geometry. In a dataset, we can find items, or nodes and links, and its attributes. Data can be static, or dynamic (e.g. a data stream) and the items/nodes attributes can be ordered or categorical.

WHY: It refers to the tasks abstraction that must feature mainly one action (a verb) and one target (a noun). Task abstraction aims to clarify what is the main purpose of a visualization, and its potential secondary purposes. It can vary from high to low level (meaning depending on how precise you want to define it), and range from presenting trends (at high level, it would be to consume data) to identifying outliers.

HOW: It refers to the design choices to visualize the data and to interact for performing the tasks. The two objectives here are to decide which visual channels like size, color, etc. will represent the data, and to choose the right marks, or the visual representations for the data (geometric primitives like lines, points, areas) for the visualization. At this stage, the idea is to choose the visual encoding and the idiom (or representation) that best suits the WHAT and WHY, to develop the visualization accordingly. This "HOW" part relies on two principles: expressiveness and effectiveness. The first one means that the chosen visual encoding should express all (and only) the information contained in the dataset, and that the visualization should show the information as it was in the dataset, in terms of data characteristics and its links with the chosen idioms. The effectiveness means that the chosen channels should always be the highest ranked ones, according to the nature of the attributes that they represent (these links between attributes and channels are referenced in the literature).

To illustrate this concept, we want to present some examples of visualizations like the ones we used further in this work. First, a bar chart (HOW) allows to summarize distribution (WHY), and to show extremes (WHY) if it also uses order (ascending or descending). Indeed, Elzer et al. [3] showed its efficiency for this kind of tasks, but note that another possible idiom for these tasks is the stacked bar chart, as Indramoto et al. [4] explained it. However, in the stack bar chart, the focus is more on combining both single-attribute and overall-attribute comparisons rather than making only single-attribute comparisons for one or more dataset (this is our case, see section 3). Furthermore, notice that here we derived the original dataset, a table, to a network dataset. In this case, following Munzner's framework, this kind of dataset is composed by nodes and links (whereas tables are composed by items) - it can be relevant to show these links or not, depending on the task. About our project data, remember that one of our aims is also to discover a new thematic classification. As Ochs et al. [5] showed it, ontologies manipulations and representations are crucial nowadays, but required much work, so they presented a software framework for doing

the following tasks: derivation, clustering and visualization as a network. So, based on their study, we can note that another visualization for ontologies is the treemap [6]. In our case, we used this last representation, and also the radial force representation (see section 3); note that in both cases, one of the most critical point is the usage of forces to separate the nodes, depending on one attribute or relationship. Hilbert et al. explained the usefulness and importance of the them in a network visualization; indeed, forces allow to separate and form groups, also called clusters [7]. This approach is useful for our work because we want to permit the public policymakers to make decisions based on visualizations that show a new classification, so in this case it could be shown thanks to the use of clustering (see Fig.**Error! Not a valid bookmark self-reference.**).

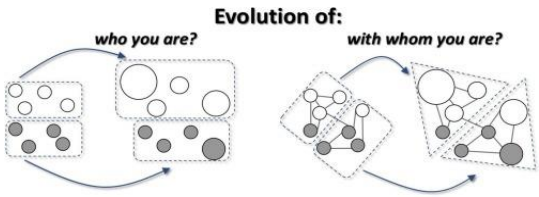


Fig. 2. Network visualization with forces for clustering by Hilbert et al.

2.2 Projects with similar issues

Here, we will present some related work that faced the same issues that we did, either from the policymakers' point of view or the visual analytics tools designers' one.

First, about public policy and data-driven policy making, Brazil had useful data about some activities in their cities, but the authorities were not using them for prioritizing the different public policies. Thus, Petrini et al. [8] applied an analytic hierarchy process (AHP) on their data and then they build some visualizations that show them adequately. As the policymakers were evaluating various priorities at the same time (environmental, economic, social), the stack bar chart was the good idiom (Fig.**Error! Not a valid bookmark self-reference.**)

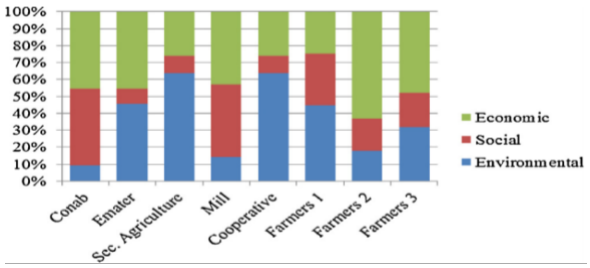
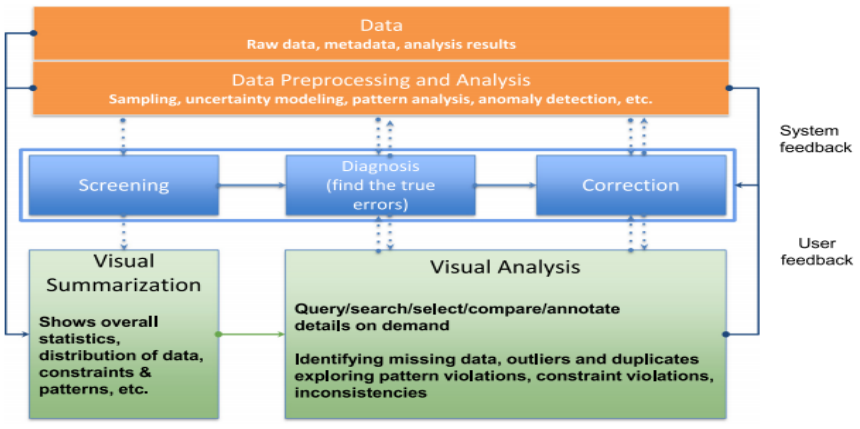


Fig. 3. Classification of thematic by priorities (multiple values for the priorities) by Petrini et al

But first, clean data and metadata is necessary: it is a common but complex issue to deal with uncleaned data. So, Liu et al. [9] proposed a framework for cleansing



(Fig.**Error! Not a valid bookmark self-reference.**)

Fig. 4. Visual analytics framework for steering data quality by Liu et al.

We can note that their process could be a complementary approach to Munzner's one, because they gave more importance to the steps of creation and evaluation of the visualization, whereas Munzner's framework focuses on abstraction (what, why, how).

Moreover, even when the ontologies have been created especially for the final users, there are real needs of availability and accuracy, meaning that these ontologies would be useless otherwise. About this issue, Kamdar et al. made a study about the usage/the access by the users of the ontologies in the biomedical field [0], about which queries the specialists made, and how they combined results.In this paper, we can see the importance of creating ontologies, that they must be user or task/issue-designed, and that they can also be viewed from a "macro" point of view, where ontologies can be combined between themselves. In the Fig.**Error! Not a valid bookmark self-reference.**, we can see how ontologies can be built.

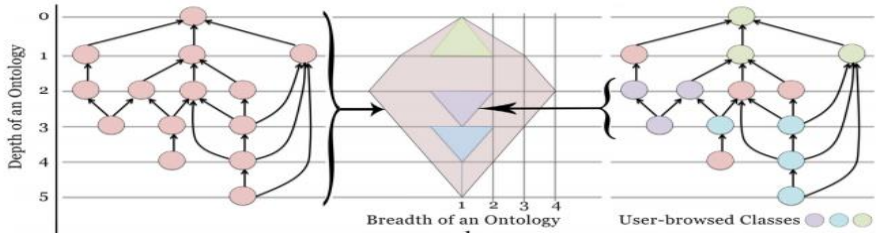


Fig. 5. Analysis of the composition of ontologies and their depth by Kamdar et al.

To sum up, in this section we have described some common issues with our project: ontologies or classifications are truly required for policy making, sometimes with an additional classification (priorities: "meta-classification", Petrini et al.). Then we have noted that other frameworks for creating visualizations than the Munzner's one exist, with other focuses than abstraction, such as cleansing (Liu et al.). We have also seen that these classifications must be accessible and accurate (Kamdar et al.). That leads us to our case study: the classification of the DANE metadata and its usage in visualizations. Currently, the DANE public policy-making tools don't satisfy their final users, because the classification used does not fit with policy making, and the visualizations are not appropriate to the tasks that the policymakers want to perform. Particularly, they need to be able to discover (identify) easily which statistical operation or administrative register is the most useful for a specific policy. Fig.**Error! Not a valid bookmark self-reference.** and Fig.**Error! Not a valid bookmark self-reference.** show examples of the current classification and visualizations. The classification may be too generic/inappropriate, and the visualizations don't show where the information is.

Fig. 6. Classification by global/macro topic (in Spanish, source: DANE website)

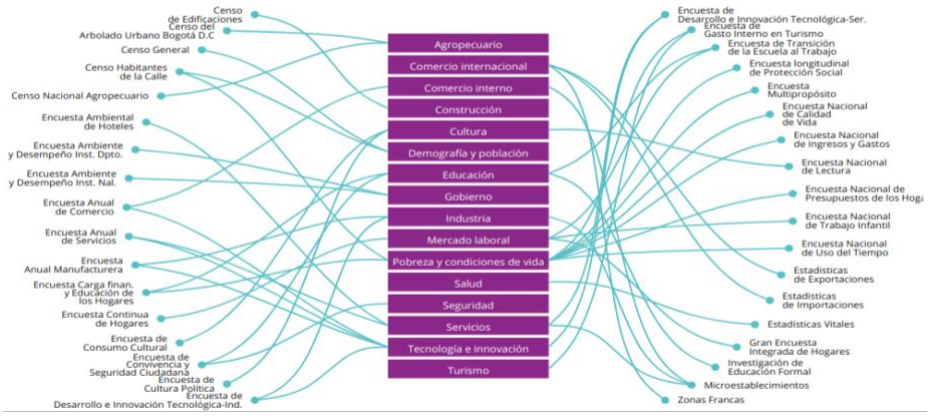


Fig. 7. Classification by sub-themes, linked to the surveys (in Spanish, source: DANE website)

3 Applying the visual analytics framework

Here we will present how we applied the framework for our three main tasks. But, before that, we should explain how we built a new classification from the metadata, resulting in a new dataset, derived from the others, so we can build task 2 (T2) and task 3 (T3) visualizations. The creation of this dataset can be considered as task 0 **(T0)**, called a derivation task. To do it, we used natural language processing on the original datasets, an acceptable way since the datasets were about 500 lines and 20 columns maximum, after cleansing. To use a commercial natural language processing tool was a valid option, but we created our own to process the words and build the new dataset in the same tool. First, it read all the lines of the file, build a dictionary of keywords that are repeated more than X times (using an exclusion list: determinants or repetitive and useless words such as register), and build the nodes and the links of the new dataset, based on the keywords occurrences in the metadata for each item. Here we explain the three abstractions, and our prototypes, obtained by applied the framework. You will find a summary for each task (details in the next sections):

T1: How many statistical operations and administrative registers are there for each topic (in the original classification)?

T2: How many statistical operations and administrative registers are there for each new topic, considering a new classification coming from the metadata?

T3: Which administrative registers and statistical operations are more related to a specific topic (new classification)?

Task 1 (T1)*What*: the original datasets – three tables, two of administrative registers and one of statistical operations (items), with, among others, the following attributes: name (categorical attribute), kind of data (categorical) and thematic area (categorical)

Why: **summarize the distribution (considering the old classification)**

How: idiom: bar chart; mark: lines; channel: depending on the visualizations, vertical or horizontal position, and color hue, one for the registers, one for the operations

Prototype: see Fig.**Error! Not a valid bookmark self-reference.** and Fig.**Error! Not a valid bookmark self-reference.**

Principal insight (see more in section 6): large difference between the numbers of administrative registers and statistical operations on the macro-category "Economics"

Task 2 (T2)

What: a new dataset derived from the previous tables – a network where the nodes (items) are the administrative registers, or the statistical operations, or the keywords of a new classification, and the links represent when an administrative register or a statistical operation matches with a keyword, one or more time; some attributes: name (categorical attribute) and new keyword groups (categorical attribute).

Why: **summarize the distribution (considering the new classification)**

How: idiom: treemap; mark: point; channel: color hue, spatial region

Prototype: see Fig.**Error! Not a valid bookmark self-reference.**, and also Fig.**Error! Not a valid bookmark self-reference.** (auxiliary visualization - details on demand)

Principal insight (see more in section 6): "Labor market" was the penultimate sub-theme in the old classification vs. "Companies" is the 4th with our new classification

Task 3 (T3)

What: a new dataset derived from the original ones (the same as in task 2)

Why: **identify features/extremes** (which nodes are **more related** to a theme, **with the new classification)**

How: idiom: radial force; mark: points; channel: radial position and color hue

Prototype: see Fig.**Error! Not a valid bookmark self-reference.**

Principal insight (more in section 6): with the keyword "Health", "Individual register of health service delivery– RIPS", is the most important administrative register.

As explained in previous sections, the main objective here is to use and understand better the DANE data to allow decision-making, resulting into two main aims: first, to determine new topics (useful for decision making on public policies) that can emerge from the metadata, and to evaluate which different statistical operations and administrative registers are more linked to a topic, and secondly, once found this

information, to provide some appropriate visualizations to the policymakers. Therefore, we developed for this case study a visual analytics tool, by applying the Munzner's framework. As there were three main different tasks, it is composed of three main components: for the task T1, several context visualizations to analyze the current state of the information held by the DANE (3.1), then for the task T2, a treemap visualization to understand the results of the natural language processing used to find more relevant major topics (a new thematic classification for decision making) around the DANE's datasets (3.2), and finally for the task T3, a radial force visualization to navigate between and into the identified topics and to provide a final tool for policymakers that shows which are the most relevant sources of information according to a topic (3.3).

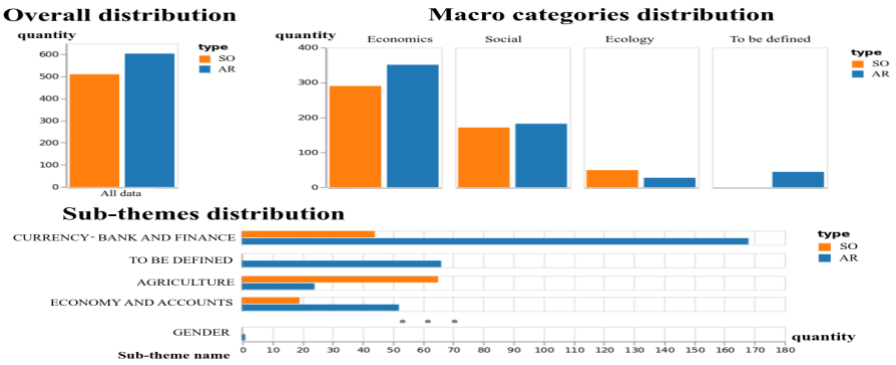
3.1 Task 1: general and contextualization task, on the original dataset

The first set of visualizations aims to represent the current inventory. The main task T1 is to *summarize* the *distributions* (WHY) of both datasets to answer the questions:

- How many statistical operations and administrative registers exist here?
- What is the proportion of statistical operations and administrative registers in the three major topics (economics, social and environmental)?
- What is the proportion of statistical operations and administrative registers in each of the 30+ specific topics (for example, health, education, etc.)?

The datasets that we used were two inventories of administrative registers and one inventory of statistical operations (WHAT) (three tables in total).

Based on the analysis made using the Munzner's framework, the best visual encoding (HOW) to provide this kind of overview is to use bar charts where the statistical operations and administrative registers are differentiated by colors. The horizontal position indicates that there are several categories (administrative registers VS statistical operations, economics VS social ...), whereas vertically the size of the bars shows items quantities differences. Fig.**Error! Not a valid bookmark self-reference.** shows the first three general and context visualizations developed. Fig.**9** shows two other visualizations developed to present the distribution of the attribute "sub-theme", allowing to know the global distribution of these sub-themes for all these administrative registers and statistical operations. We used the same encoding since it is still the best one for *summarizing the distributions*, and to *identify extremes* (secondary task), we used the technique "separate order and align" for that purpose). About this last point, if considering only administrative registers (on the left in Fig 9.), the most present sub-theme is "currency – bank and finance" whereas looking



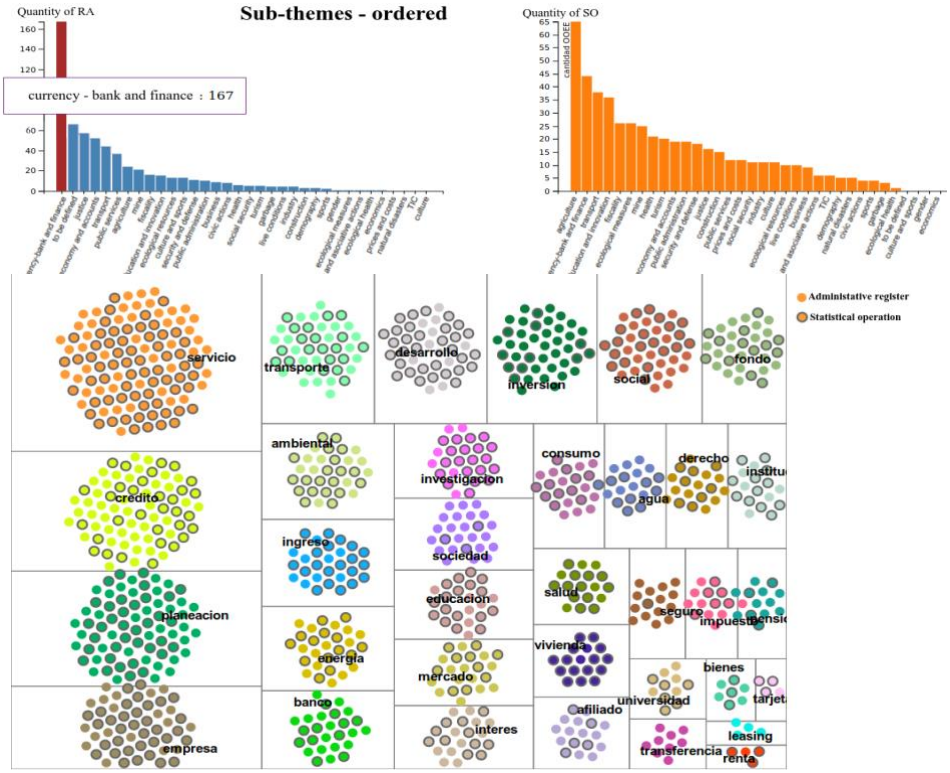
only at statistical operations (on the right), it is “agriculture”.

Fig. 8. Bar chart visualizations based on sub-themes from the original data.

Fig. 9. Bar chart visualizations about original sub-themes, with “separate order and align”.

3.2 Task 2: new dataset, new classification, but which distribution?

Here we present the T2 task visualization: *summarize the distribution with the new classification* (WHY). We used a treemap and a bar chart (HOW). It uses the derived dataset, that contains nodes and links (WHAT), each node being a statistical operation or an administrative register, and each link being a relationship between nodes, particularly between a “register/operation” node and a “keyword” node. We used clustering for grouping them by (new) themes. For separating the clusters, we used the force-in-a-box algorithm (https://github.com/john-guerra/forceInABox), by J. Guerra. Thus, this visualization allows to have a global vision (*summarize*



distribution) of the new different themes: transport (*transporte*), research (*investigación*) etc. It also allows to detect that services (*servicio*) and credits (*crédito*) are the themes that most appear (*identify extremes*). Both in the tree map (by nature) and in the bar chart (common technique), we applied the technique “separate order and align” for completing this secondary task. In Fig.**Error! Not a valid bookmark self-reference.**, administrative registers and statistical operations are considered,

whereas in Fig.***12***, they are separated. However, we notice that the top themes are globally the same, considering them separately (Fig.***12***: services (*servicio*), planning (*planeación*)) or in total (Fig.**10**: top3:services, credits, planning).

Fig. 10. Tree map chart visualization on new themes (derived data – DANE data in Spanish).

The following visualization is a table, coupled to the previous treemap that give information about an item by clicking on it in the treemap, to get specific information.

Table for detailed information

Name: ESTADISTICAS DE SERVICIOS AEROPORTUARIOS
Type: OOEE
Statistical operations related to:
Operations: Consolidar informacion estadistica de los servicios de escala en a eropuerto (Handling)

Fig. 11. Auxiliary view, a table, of the treemap visualization (DANE data in Spanish)

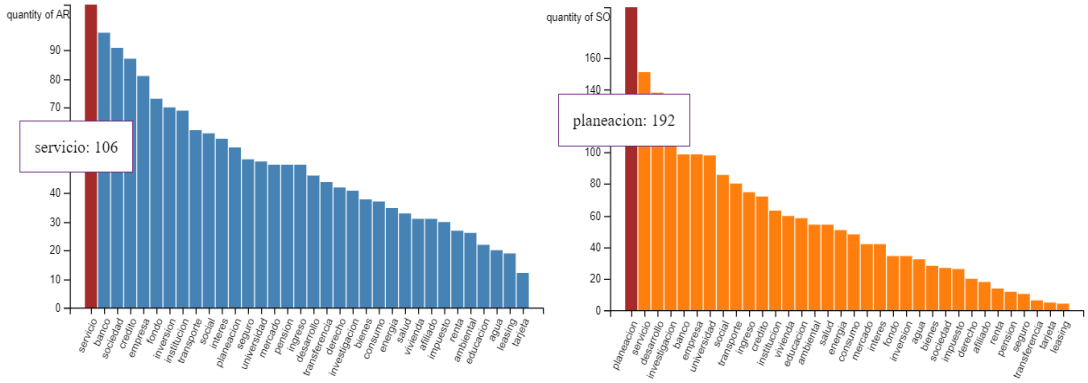
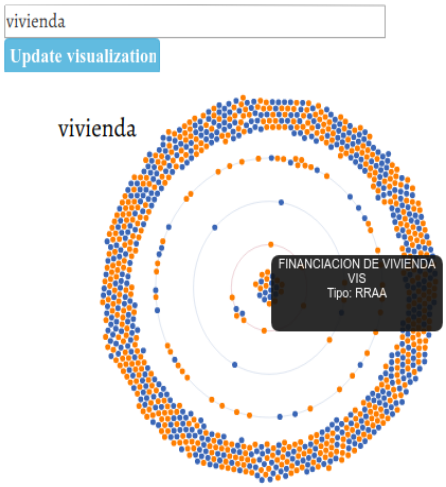


Fig. 12. Bar chart visualization on new themes, in blue the administrative registers, in orange the statistical operations (derived data – DANE data in Spanish).

3.3 Task 3: given a keyword, which are the items more linked to it?

To navigate between the new topics and the nodes associated to it, we created a visualization where the user can type a word, and then explore the statistical operations and administrative registers that feature this keyword in their metadata, in



Keywords:
credito, servicio, inversion, fondo, empresa, banco, planeacion, social, ingreso, interes, transporte, desarrollo, pension, consumo, institucion, investigacion, vivienda, transferencia, tarjeta, derecho, sociedad, leasing

This table shows in top down order the administrative registers and statistical operations that are related with the chosen keyword

Tipo	Nombre del Nodo	Puntaje
RRAA	FINANCIACION DE VIVIENDA VIS	14
OOEE	ESTADISTICAS SOBRE VIVIENDA DE INTERES PRIORITARIO Y SOCIAL INICIADAS CON APOYO DE FONVIVIENDA	13
RRAA	INFORME SEMANAL PRINCIPALES CUENTAS ACTIVAS Y PASIVAS - SALDOS AL CIERRE	11
RRAA	FORMATO DE CUENTAS NO PUC PARA EL CALCULO DEL PATRIMONIO ADECUADO	9
RRAA	INFORMACION COSECHAS CREDITOS DE VIVIENDA	9
OOEE	CENSO DE POBLACION Y VIVIENDA (CNPV)	9
OOEE	INDICE DE COSTOS DE LA CONSTRUCCION DE VIVIENDA (ICCV)	9
RRAA	FORMATO DE CUENTAS NO PUC PARA EL CALCULO DEL PATRIMONIO ADECUADO SECCIONES ESPECIALIZADAS DE AHORRO Y CREDITO DE LAS CAJAS DE COMPENSACION FAMILIAR	7
RRAA	REPORTE DE AVANCE FISICO ACUMULADO DE EJECUCION DE MISN	7
RRAA	INFORMACION MONTOS Y NUMERO DE CREDITOS POR COSECHAS	6
OOEE	POTENCIAL DE VIVIENDA EN LOS MACROPROYECTOS DE INTERES SOCIAL NACIONAL (MISN) ADOPTADOS	6
OOEE	VIVIENDAS CONSTRUIDAS EN LOS MACRO PROYECTOS DE INTERES SOCIAL NACIONAL (MISN) ADOPTADOS	6

top-down order. As a result, the main task T3 of the visualization is to *identify features* (WHY).

Fig. 13. Radial force visualization and its auxiliary view on the right (DANE data in Spanish)

In this visualization, we used the radial force idiom (HOW), using forces to order the items depending on how much they fit with the written keyword (separate and order). It means that, after choosing the keyword, the user can see that the statistical operations and administrative registers that contain the keyword are more or less attracted to the center depending on the number of occurrences (the ones that do not contain the keyword at all remain in the border), allowing to identify the extremes (similar items can be identified by the spatial region where they are). Fig.**1.** DANE's data distribution and explanation of the problematic: does a new thematic classification coming from the metadata of the administrative registers and statistic operation exist?

Based on these considerations, the main objective of this project is to build a tool to understand and visualize the DANE data, particularly by organizing it through the topics and keywords present among the metadata of different groups of statistical operations and administrative registers, ultimately allowing decision-makers to have right overviews of topics and to find which statistical operations and administrative registers are related to one in particular, thanks to this classification (see Fig.**Error! Not a valid bookmark self-reference.**).

4 Related work

4.1 Tamara Munzner's framework

For this project, we used Munzner's visualization framework [2] to abstract and understand the data, the users' tasks and to choose the best idioms that allow the users to complete these tasks. It has three dimensions: the WHAT, the WHY and the HOW.

WHAT: It refers to the available data for the visualization. The basic abstractions of the dataset arrangements are tables, networks, fields and geometry. In a dataset, we can find items, or nodes and links, and its attributes. Data can be static, or dynamic (e.g. a data stream) and the items/nodes attributes can be ordered or categorical.

WHY: It refers to the tasks abstraction that must feature mainly one action (a verb) and one target (a noun). Task abstraction aims to clarify what is the main purpose of a visualization, and its potential secondary purposes. It can vary from high to low level (meaning depending on how precise you want to define it), and range from presenting trends (at high level, it would be to consume data) to identifying outliers.

HOW: It refers to the design choices to visualize the data and to interact for performing the tasks. The two objectives here are to decide which visual channels like size, color, etc. will represent the data, and to choose the right marks, or the visual representations for the data (geometric primitives like lines, points, areas) for the visualization. At this stage, the idea is to choose the visual encoding and the idiom (or representation) that best suits the WHAT and WHY, to develop the visualization accordingly. This "HOW" part relies on two principles: expressiveness and

effectiveness. The first one means that the chosen visual encoding should express all (and only) the information contained in the dataset, and that the visualization should show the information as it was in the dataset, in terms of data characteristics and its links with the chosen idioms. The effectiveness means that the chosen channels should always be the highest ranked ones, according to the nature of the attributes that they represent (these links between attributes and channels are referenced in the literature).

To illustrate this concept, we want to present some examples of visualizations like the ones we used further in this work. First, a bar chart (HOW) allows to summarize distribution (WHY), and to show extremes (WHY) if it also uses order (ascending or descending). Indeed, Elzer et al. [3] showed its efficiency for this kind of tasks, but note that another possible idiom for these tasks is the stacked bar chart, as Indramoto et al. [4] explained it. However, in the stack bar chart, the focus is more on combining both single-attribute and overall-attribute comparisons rather than making only single-attribute comparisons for one or more dataset (this is our case, see section 3). Furthermore, notice that here we derived the original dataset, a table, to a network dataset. In this case, following Munzner's framework, this kind of dataset is composed by nodes and links (whereas tables are composed by items) - it can be relevant to show these links or not, depending on the task. About our project data, remember that one of our aims is also to discover a new thematic classification. As Ochs et al. [5] showed it, ontologies manipulations and representations are crucial nowadays, but required much work, so they presented a software framework for doing the following tasks: derivation, clustering and visualization as a network. So, based on their study, we can note that another visualization for ontologies is the treemap [6]. In our case, we used this last representation, and also the radial force representation (see section 3); note that in both cases, one of the most critical point is the usage of forces to separate the nodes, depending on one attribute or relationship. Hilbert et al. explained the usefulness and importance of the them in a network visualization; indeed, forces allow to separate and form groups, also called clusters [7]. This approach is useful for our work because we want to permit the public policymakers to make decisions based on visualizations that show a new classification, so in this case it could be shown thanks to the use of clustering (see Fig.**Error! Not a valid bookmark self-reference.**).

Fig. 2. Network visualization with forces for clustering by Hilbert et al.

4.2 Projects with similar issues

Here, we will present some related work that faced the same issues that we did, either from the policymakers' point of view or the visual analytics tools designers' one.

First, about public policy and data-driven policy making, Brazil had useful data about some activities in their cities, but the authorities were not using them for prioritizing the different public policies. Thus, Petrini et al. [8] applied an analytic hierarchy process (AHP) on their data and then they build some visualizations that

show them adequately. As the policymakers were evaluating various priorities at the same time (environmental, economic, social), the stack bar chart was the good idiom (Fig.**Error! Not a valid bookmark self-reference.**)

Fig. 3. Classification of thematic by priorities (multiple values for the priorities) by Petrini et al

But first, clean data and metadata is necessary: it is a common but complex issue to deal with uncleaned data. So, Liu et al. [9] proposed a framework for cleansing (Fig.**Error! Not a valid bookmark self-reference.**)

Fig. 4. Visual analytics framework for steering data quality by Liu et al.

We can note that their process could be a complementary approach to Munzner's one, because they gave more importance to the steps of creation and evaluation of the visualization, whereas Munzner's framework focuses on abstraction (what, why, how).

Moreover, even when the ontologies have been created especially for the final users, there are real needs of availability and accuracy, meaning that these ontologies would be useless otherwise. About this issue, Kamdar et al. made a study about the usage/the access by the users of the ontologies in the biomedical field [0], about which queries the specialists made, and how they combined results.In this paper, we can see the importance of creating ontologies, that they must be user or task/issue-designed, and that they can also be viewed from a "macro" point of view, where ontologies can be combined between themselves. In the Fig.**Error! Not a valid bookmark self-reference.**, we can see how ontologies can be built.

Fig. 5. Analysis of the composition of ontologies and their depth by Kamdar et al.

To sum up, in this section we have described some common issues with our project: ontologies or classifications are truly required for policy making, sometimes with an additional classification (priorities: "meta-classification", Petrini et al.). Then we have noted that other frameworks for creating visualizations than the Munzner's one exist, with other focuses than abstraction, such as cleansing (Liu et al.). We have also seen that these classifications must be accessible and accurate (Kamdar et al.). That leads us to our case study: the classification of the DANE metadata and its usage in visualizations. Currently, the DANE public policy-making tools don't satisfy their final users, because the classification used does not fit with policy making, and the visualizations are not appropriate to the tasks that the policymakers want to perform. Particularly, they need to be able to discover (identify) easily which statistical operation or administrative register is the most useful for a specific policy. Fig.**Error! Not a valid bookmark self-reference.** and Fig.**Error! Not a valid bookmark self-reference.** show examples of the current classification and visualizations. The

classification may be too generic/inappropriate, and the visualizations don't show where the information is.

Fig. 6. Classification by global/macro topic (in Spanish, source: DANE website)

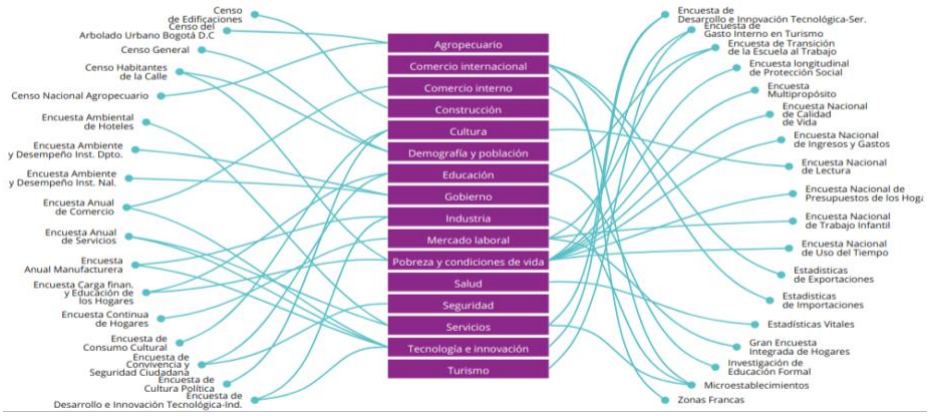


Fig. 7. Classification by sub-themes, linked to the surveys (in Spanish, source: DANE website)

5 Applying the visual analytics framework

Here we will present how we applied the framework for our three main tasks. But, before that, we should explain how we built a new classification from the metadata, resulting in a new dataset, derived from the others, so we can build task 2 (T2) and task 3 (T3) visualizations. The creation of this dataset can be considered as task 0 **(T0)**, called a derivation task. To do it, we used natural language processing on the original datasets, an acceptable way since the datasets were about 500 lines and 20 columns maximum, after cleansing. To use a commercial natural language processing tool was a valid option, but we created our own to process the words and build the new dataset in the same tool. First, it read all the lines of the file, build a dictionary of keywords that are repeated more than X times (using an exclusion list: determinants or repetitive and useless words such as register), and build the nodes and the links of the new dataset, based on the keywords occurrences in the metadata for each item. Here we explain the three abstractions, and our prototypes, obtained by applied the framework. You will find a summary for each task (details in the next sections):

T1: How many statistical operations and administrative registers are there for each topic (in the original classification)?

T2: How many statistical operations and administrative registers are there for each new topic, considering a new classification coming from the metadata?

T3: Which administrative registers and statistical operations are more related to a specific topic (new classification)?

Task 1 (T1)*What*: the original datasets – three tables, two of administrative registers and one of statistical operations (items), with, among others, the following attributes: name (categorical attribute), kind of data (categorical) and thematic area (categorical)

Why: **summarize the distribution (considering the old classification)**

How: idiom: bar chart; mark: lines; channel: depending on the visualizations, vertical or horizontal position, and color hue, one for the registers, one for the operations

Prototype: see Fig.**Error! Not a valid bookmark self-reference.** and Fig.**Error! Not a valid bookmark self-reference.**

Principal insight (see more in section 6): large difference between the numbers of administrative registers and statistical operations on the macro-category "Economics"

Task 2 (T2)

What: a new dataset derived from the previous tables – a network where the nodes (items) are the administrative registers, or the statistical operations, or the keywords of a new classification, and the links represent when an administrative register or a statistical operation matches with a keyword, one or more time; some attributes: name (categorical attribute) and new keyword groups (categorical attribute).

Why: **summarize the distribution (considering the new classification)**

How: idiom: treemap; mark: point; channel: color hue, spatial region

Prototype: see Fig.**Error! Not a valid bookmark self-reference.**, and also Fig.**Error! Not a valid bookmark self-reference.** (auxiliary visualization - details on demand)

Principal insight (see more in section 6): "Labor market" was the penultimate sub-theme in the old classification vs. "Companies" is the 4th with our new classification

Task 3 (T3)

What: a new dataset derived from the original ones (the same as in task 2)

Why: **identify features/extremes** (which nodes are **more related** to a theme, **with the new classification)**

How: idiom: radial force; mark: points; channel: radial position and color hue

Prototype: see Fig.**Error! Not a valid bookmark self-reference.**

Principal insight (more in section 6): with the keyword "Health", "Individual register of health service delivery– RIPS", is the most important administrative register.

As explained in previous sections, the main objective here is to use and understand better the DANE data to allow decision-making, resulting into two main aims: first, to determine new topics (useful for decision making on public policies) that can emerge from the metadata, and to evaluate which different statistical operations and administrative registers are more linked to a topic, and secondly, once found this information, to provide some appropriate visualizations to the policymakers. Therefore, we developed for this case study a visual analytics tool, by applying the Munzner's framework. As there were three main different tasks, it is composed of three main components: for the task T1, several context visualizations to analyze the current state of the information held by the DANE (3.1), then for the task T2, a treemap visualization to understand the results of the natural language processing used to find more relevant major topics (a new thematic classification for decision making) around the DANE's datasets (3.2), and finally for the task T3, a radial force visualization to navigate between and into the identified topics and to provide a final tool for policymakers that shows which are the most relevant sources of information according to a topic (3.3).

5.1 Task 1: general and contextualization task, on the original dataset

The first set of visualizations aims to represent the current inventory. The main task T1 is to *summarize* the *distributions* (WHY) of both datasets to answer the questions:

- How many statistical operations and administrative registers exist here?
- What is the proportion of statistical operations and administrative registers in the three major topics (economics, social and environmental)?
- What is the proportion of statistical operations and administrative registers in each of the 30+ specific topics (for example, health, education, etc.)?

The datasets that we used were two inventories of administrative registers and one inventory of statistical operations (WHAT) (three tables in total).

Based on the analysis made using the Munzner's framework, the best visual encoding (HOW) to provide this kind of overview is to use bar charts where the statistical operations and administrative registers are differentiated by colors. The horizontal position indicates that there are several categories (administrative registers VS statistical operations, economics VS social ...), whereas vertically the size of the bars shows items quantities differences. Fig.**Error! Not a valid bookmark self-reference.** shows the first three general and context visualizations developed. Fig.**9** shows two other visualizations developed to present the distribution of the attribute "sub-theme", allowing to know the global distribution of these sub-themes for all these administrative registers and statistical operations. We used the same encoding since it is still the best one for *summarizing the distributions*, and to *identify extremes* (secondary task), we used the technique "separate order and align" for that purpose). About this last point, if considering only administrative registers (on the left in Fig 9.), the most present sub-theme is "currency – bank and finance" whereas looking only at statistical operations (on the right), it is "agriculture".

Fig. 8. Bar chart visualizations based on sub-themes from the original data.

Fig. 9. Bar chart visualizations about original sub-themes, with "separate order and align".

5.2 Task 2: new dataset, new classification, but which distribution?

Here we present the T2 task visualization: *summarize the distribution with the new classification* (WHY). We used a treemap and a bar chart (HOW). It uses the derived dataset, that contains nodes and links (WHAT), each node being a statistical operation or an administrative register, and each link being a relationship between nodes, particularly between a "register/operation" node and a "keyword" node. We used clustering for grouping them by (new) themes. For separating the clusters, we used the force-in-a-box algorithm (https://github.com/john-guerra/forceInABox), by J. Guerra. Thus, this visualization allows to have a global vision (*summarize distribution*) of the new different themes: transport (*transporte*), research (*investigación*) etc. It also allows to detect that services (*servicio*) and credits (*crédito*) are the themes that most appear (*identify extremes*). Both in the tree map (by nature) and in the bar chart (common technique), we applied the technique "separate order and align" for completing this secondary task. In Fig.**Error! Not a valid bookmark self-reference.**, administrative registers and statistical operations are considered, whereas in Fig.***12***, they are separated. However, we notice that the top themes are globally the same, considering them separately (Fig.***12***: services (*servicio*), planning (*planeación*)) or in total (Fig.**10**: top3:services, credits, planning).

Fig. 10. Tree map chart visualization on new themes (derived data – DANE data in Spanish).

The following visualization is a table, coupled to the previous treemap that give information about an item by clicking on it in the treemap, to get specific information.

Fig. 11. Auxiliary view, a table, of the treemap visualization (DANE data in Spanish)

Fig. 12. Bar chart visualization on new themes, in blue the administrative registers, in orange the statistical operations (derived data – DANE data in Spanish).

5.3 Task 3: given a keyword, which are the items more linked to it?

To navigate between the new topics and the nodes associated to it, we created a visualization where the user can type a word, and then explore the statistical operations and administrative registers that feature this keyword in their metadata, in top-down order. As a result, the main task T3 of the visualization is to *identify features* (WHY).

Fig. 13 shows this visualization, where the statistical operations are orange and the administrative registers are blue (categorical attribute of the items, so using color hue is effective). In addition, by putting the mouse over any item, the user can see the item name and its type.

Moreover, to help *identifying the extremes* (WHY), right to the visualization, we added an auxiliary view, a table, where the elements are ordered in descendant order,

so the user can find the statistical operation or the administrative register that features the major occurrences of the keyword, considering all its attributes, in its metadata.

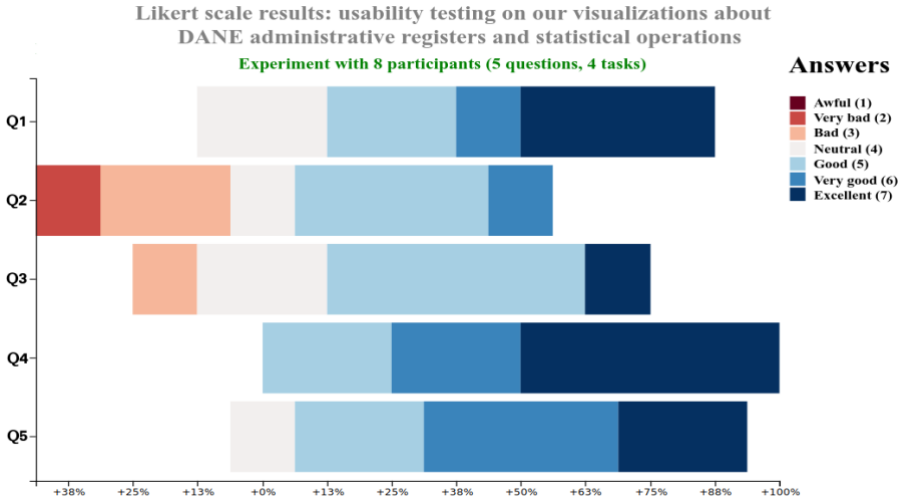
6 Experiment and results

6.1 Experiment

To validate our work, we organized an experiment where the experts from the DANE were invited to try our tool with all the visualizations created according to the tasks explained before, and according to our application of the visual analytics framework. In total, there were 8 participants, 6 females and 2 males. 5 of them were working in the R&D department (in other words, “our clients”, the people who asked for the tool) and 3 were working in the department responsible for the planning based on statistics (in other words, the final users of our future tool, apart from the policymakers). All the participants had to follow these instructions: “first, try to get new themes about registers and operations (with the treemap now, but actually, during the experiment, with a network visualization, see section 6.2); then get some information about one item (with the coupled table); after that, write a word about one theme of your interest and discover which are the registers and operations more related with this keyword (with the radial visualization); finally, read the name of the most important register/operation in the coupled table”.

6.2 Results

First, note that in this experience, the users were using and evaluating a previous version of our visual analytics tool that the one in this paper. Thanks to these results, we have been able to correct our visualizations. So, these results show how our work has evolved. It also shows that to apply the visual analytics framework may often require iterations, with several users experiments (and that some tasks may require more iterations than others). The following results come from a post-experiment questionnaire, where the users evaluated the quality of our visualizations (usability and completion of the tasks). Using Likert scale, we asked closed questions, one for each visualization and task. The idea of this usability testing questionnaire was to determine if the visualizations were answering or not to the previously defined issues



of the DANE (with the Visual Analytics Framework, visualizations must be built according to the tasks defined using abstraction, from the original tasks described by users). Moreover, we asked open questions to get some feedback, for making corrections. Finally, according to the results, that would be discussed in the next section, we created the final visualizations that are presented in the paper in section 3. So notice that in this experience, the tool had only two parts not three (the context visualizations part, for the task T1, were missing), composed of four visualizations: a network visualization (where the forces were not separating the clusters of nodes as good as in the treemap, which is the "evolution" of this visualization after the experience) and its coupled table, for the task T2, and the radial force visualization and its coupled table, for the task T3.

Fig. 14. Results of the experiment with the users (Likert scale graph).

- Q1- What is your general impression with this tool?
- Q2- Have you been able to explore a new classification of the items?
- Q3- Have you been able to obtain the detail for one of these items?
- Q4- Have you been able to discover the items in relation with a theme?
- Q5- Have you been able to identify the item more related with a theme?

7 Discussion

According to the results, both the quality results shown on the Likert scale graph (see Fig.**14**), and the feedback given by the experts indicate that the task they performed with more difficulty (almost 50% of the participants grade it between 1 and 4 included) was the one asked in Q2: explore the new classification in the network visualization (T2). On the contrary, the easiest task for them (100% of them grade it between 4 and 7 included) was the one asked in Q4: discover the items in relation with one theme in the radial force visualization (T3). As a result, we did more corrections on the visualization used in Q2, so it is the only visualization where we had to completely modify the idiom (visual encoding) used, transforming the visualization from a network visualization at that moment (clusters might not be appearing so clearly, but the relationships/links between elements do appear better) to a treemap now (on the contrary, the focus is clearly on clustering) – to sum up, both idioms use nodes and forces, but, in our case, the clustering was clearer with the treemap visualization.

Additionally, we noticed, both in the comments and the qualitative results (apart from Q2 results), thanks to Q1 results, with 25% of neutral grade (4), that something might be missing, apart from the current visualizations and their corrections, something that shows better the purpose of our tool and why did we create it. In other terms: to understand where we are going, we should know what were the original data? Were there some insights in the original data? The next visualizations are showing different data? What is the difference with the new thematic classification? What are the new insights? All these questions can be considered as a first task for the users (T1). So, that is why we finally added the context visualizations part and organized the tool into three parts and not two: context - T1 (original data), treemap - T2 (derived data) and radial force visualization - T3 (derived data).

8 Insights

The new (corrected) visualizations (presented in the section 3) allow to discover some insights about the statistical data and their classification in new themes.

First, thanks to the context visualizations based on the initial data, we discovered than globally more information could be generated because there are more registers than operations (remember that the administrative registers create the statistical operations). We noticed this particularly for the macro category “Economics”. By looking also at the visualization by sub-themes, it is confirmed, with more details: the sub-themes “Currency, banks and finance” and “Accounts and economics” appear as some of the ones where there is much difference between the number of registers and the number of operations, and both are belonging to the macro category “Economics”.

Then, thanks to the new derived data and the treemap visualization, we can discover other insights. First, by looking to the terms that appear, on one hand this visualization confirms that the macro category and sub-theme currently used are quite coherent with the metadata. For example, we can find as new keywords “research”

(*investigación*), "credits" (*crédito*), "market" (*mercado*), whereas that, in the current sub-theme, there were "education and innovation", "business", "economy and accounts" etc. But, in the other hand, this visualization also suggests that the categories and sub-themes used currently may not be so accurate in term of distribution, because in the current classification, "labor market" is the penultimate sub-theme whereas "company" (*empresa*) is the fourth one with our new thematic classification. Another insight is that this visualization gives us more details about a previous insight from the first visualizations: in the current category "Economics", the focus for generating new operations from registers should be done more precisely on the ones that contains in their metadata the words "leasing" or "transactions", since we observe in this visualization that there are no statistical operations about these subjects, only registers.

Finally, some other insights have been revealed thanks to the radial force visualization. As our final tool in this study for decision making on public policies, we can notice that, thanks to it, people, without being an expert about data manipulation, can identify easily which registers and operations are more related to a specific theme: for example about "Housing" (the chosen keyword), the most important administrative register is "Housing financing VIS", or for "Health it is "Individual register of health service delivery – RIPS". To conclude, this last visualization, by grouping in the center the elements with more relevance but separating them into two categories, for being administrative registers or statistical operations by using two different colors, allows at a glance to notice for one thematic if the relevant elements are mostly of one kind of information (because visually it will be mostly of one color), and as a result, it can confirm that more statistical operations should be generated in this thematic or not (so we could make some "priorities" about statistical operation generations). So, for example, following with the theme "Economics", we can notice for the keyword "credits" that more statistical operations could be generated (there are much more administrative registers).

9 Conclusion

Finally, by applying our approach, we obtained useful visualizations that confirmed that the DANE owns highly relevant information for the country and that they should continue developing more data analysis tools, to provide them to public policymakers to maximize the usage of their data. The visual analytics tools permit both policymakers and citizens to locate where the relevant information is, and allows its understanding, ultimately enhancing policies and fostering data-driven businesses. So, we might think that our contribution helped the DANE to understand that their data are very valuable, and that with such approaches, these data can be classified and presented in a way that allow the policymakers to use it for public policies making. As written in the previous section, thanks to this study and its visualizations, the DANE learned, among other insights, that, even if they held a large amount of data about economics, they could explore and use it better: more statistical operations could be generated, especially about topics such as leasing or transactions. This

insight might be the most relevant because parts of it appeared in most of the different visualizations.

About future work, it could be interesting to explore other possibilities about our natural language processing tool (the one for getting the keywords that appear in the metadata), and also to think about how could we grade differently the administrative registers or statistical operations that belong to a theme (currently it is based on the number of occurrences of the keywords in the metadata – the DANE has already confirmed it interest for this avenue). Finally, another possibility of future work could be to study which visualization would be appropriate for showing the relationships between the administrative registers and the statistic operations that are linked (because this register produces this operation) while the visualization shows the clusters of nodes by thematic.

References

1. OECD, OECD Digital Government Studies Digital Government Review of Colombia Towards a Citizen-Driven Public Sector, OECD Publishing, France (2018). https://doi.org/10.1787/9789264291867-en
2. T. Munzner, Visualization Analysis and Design. 1st edn. A K Peters Visualization Series/CRC Press, Natick, Massachusetts, USA (2014). ISBN-10: 9781466508910, ISBN-13: 978-1466508910
3. S. Elzer, S. Carberry, I. Zukerman, The automated understanding of simple bar charts, Artificial Intelligence 175(2), 526–555 (2011). https://doi.org/10.1016/j.artint.2010.10.003
4. Indratmo, L. Howorko, J. Boedianto, B. Daniel, The efficacy of stacked bar charts in supporting single-attribute and overall-attribute comparisons, Visual Informatics 2(3), 155–165 (2018). https://doi.org/10.1016/j.visinf.2018.09.002
5. C. Ochs, J. Geller, Y. Perl, M.A. Musen, A unified software framework for deriving, visualizing, and exploring abstraction networks for ontologies, Journal of Biomedical Informatics 62, 90–105 (2016). https://doi.org/10.1016/j.jbi.2016.06.008
6. B. Shneiderman, Tree visualization with tree-maps: 2-d space-filling approach. ACM Transactions on Graphics, 11(1), 92–99 (1992). https://doi.org/10.1145/102377.115768
7. M. Hilbert, P. Oh, P. Monge, Evolution of what? A network approach for the detection of evolutionary forces, Social Networks 47, 38–46 (2016). https://doi.org/10.1016/j.socnet.2016.04.003
8. M.A. Petrini, J. Rocha, J.C. Brown, R. Bispo, Using an analytic hierarchy process approach to prioritize public policies addressing family farming in Brazil, Land Use Policy, 51, 85-94 (2016). https://doi.org/10.1016/j.landusepol.2015.10.029
9. S. Liu, G. Andrienko, Y. Wu, N. Cao, L. Jiang, C. Shi,Y.S Wang, S. Hong, Steering data quality with visual analytics: The complexity challenge, Visual Informatics, 2(4), 191-197 (2018). https://doi.org/10.1016/j.visinf.2018.12.001
10. M.R. Kamdar, S. Walk, T. Tudorache, M.A. Musen, Analyzing user interactions with biomedical ontologies: A visual perspective, Journal of Web Semantics, 49, 16-30 (2018). https://doi.org/10.1016/j.websem.2017.12.002