ENTREGA DE INFORMES Y/O PRODUCTOS

CONTRATO Nº 457-2017

X	DNP
	FNR
	DNPOR
	DNP MÍNIMA CUANTÍA
	DNPSGR

Nombre del contratista	Mario Andres Morales Moreno		
Supervisor	Iván Antonio Mantilla Gaviria (1) Carlos Alberto Ferrer Infante (2)		
Dependencia	Oficina de Informática		
Fecha de Recibo del Supervisor	31 de oct de 17	**Firma del Supervisor (1)**	
Fecha de Recibo del Supervisor	31 de oct de 17	**Firma del Supervisor (2)**	

Informe Nº 9 de 12

Producto Nº 1 de 1

Folios 33

Anexos

OBSERVACIONES

DNP
Rad No.: 20176630575232
Fecha: 2017-10-31 08:47:44 Usr. Rad.: VPENARANDA3
Asunto: INFORME No 9 PRODUCTO No 1 CONTRATO DNP-457-2017 MAF
Destino: GRUPO DE CORRESPONDENCIA

CONTRATO DE SERVICIOS DNP-457-2017

INFORME DE ACTIVIDADES

Del 1 al 31 de octubre de 2017

Mario Andres Morales Moreno

PARA: Iván Antonio Mantilla Gaviria
Carlos Alberto Ferrer Infante

REF: Informe de Actividades – Contrato de Servicios DNP-457-2017

FECHA: 31 de octubre de 2017

En atención a lo dispuesto por la Cláusula Octava del contrato de la referencia, de manera atenta me dirijo a usted para presentarle el informe de las actividades adelantadas durante el período comprendido entre **el 1 al 31 de octubre de 2017,** en desarrollo de dicho contrato, el cual tuvo inicio el 16 de febrero de 2017 y finaliza el 31 de diciembre de 2017.

Objeto del contrato:

Apoyar a la Oficina de Informática del Departamento Nacional de Planeación – DNP, en las actividades asociadas con la generación de insumos para realizar analítica de datos y sus componentes relacionados, que apoyen iniciativas, pilotos y soluciones en la entidad.

Actividades realizadas:

De acuerdo con la cláusula séptima del contrato a continuación se describen las actividades realizadas:

1. Apoyar a la Oficina de informática del DNP, en la coordinación, gestión de datos, ciencia de datos y en general el ciclo de vida de los datos de la entidad.

    - Continue la Definición del proyecto de IDEAM para mejorar el sistema SIRH a ser desarrollado en el segundo semestre de 2017.
    - Continue la Definición Definición del proyecto big data con el Ministerio de Salud, ADRES y la Subdirección de salud del DNP.
    - Continue el desarrollo del cronograma y Planteamiento del problema de la primera fase de trabajo en conjunto con la Dirección de Desarrollo Urbano para los proyectos de analítica de datos formulados para CAOBA, en el análisis de consumo de agua.
    - Participe en Mesa técnica con la DSEPP para definir el desarrollo del proyecto ODS y asignar responsabilidades y tareas por parte del equipo de científicos de datos para la hoja de ruta del proyecto: "Midiendo el impacto en los trinos de los políticos hombres y mujeres de Colombia".

2. Apoyar a la Oficina de Informática, en las iniciativas y soluciones desarrollados por las dependencias que impliquen la implementación de componentes de análisis de datos y que apoyen la toma de decisiones de nivel estratégico y operativo.

   - Organize y defini con la Oficina de Informática los lineamentos necesarios por el equipo de científicos de datos en temas relacionados con infraestructura y almacenamiento de datos de futuros proyectos en el DNP para el año 2018.

3. Apoyar a la Oficina de Informática, en la elaboración de informes estadísticos sobre correlación de variables, analítica descriptiva y predictiva; de fuentes de información internas y/o externas de la entidad que apoyen la toma de decisiones estratégicas.

   - Apoye técnicamente al equipo de trabajo del proyecto "Smart Regulation" identificando mejores prácticas en el uso del big data en otros paises y su aporte al sector público.
   - Presente Resultados del Grupo de Big Data del DNP, definición de anteproyectos para el 2018 y requerimientos de Pasantes ante del Dr Ivan Mantilla.

4. Apoyar a la Oficina de Informática, en la elaboración de insumos matemáticos y estadísticos que soporten los informes estadísticos solicitados por la Oficina de Informática o por las dependencias de la entidad; utilizando las herramientas de software adecuadas.

   - Ejecute el procesamiento y limpieza de documentación via OCR para el proyecto de mejora normativa.

5. Apoyar a la Oficina de Informática, en la consolidación de datos provenientes de fuentes internas y/o externas de la Entidad para facilitar las tareas de análisis.

   - Análize de información inicial proveniente de WAZE para proyectos de movilidad externos (CAOBA) e internos.

6. Apoyar a la Oficina de Informática, en la gestión de información, resultante de procesos y procedimientos de captura, almacenamiento, procesamiento y análisis de datos.

7. Apoyar a la Oficina de Informática, en la administración, configuración, programación y operación de herramientas de apoyo para la gestión de información asociada análisis de datos.

8. Apoyar a la Oficina de Informática, en las transferencias de conocimiento dirigidas a las áreas del DNP, relacionadas con las implementaciones realizadas y en los casos que se requiera.

9. Integrar Comités Asesores-Evaluadores asignados por el supervisor del contrato en relación con los procesos de contratación que requiera la Oficina de Informática del DNP.

   - Presente resultados del proyecto "Herramienta de Búsqueda de Datos Masivos (Big Data) para Mejora Normativa" en el evento "Transformando el Estado colombiano con Big Data" efectuado el día 18 de octubre de 2017.
   - Asistí virtualmente a Conferencia con CCP. WAZE Summit NY octubre 2017.

10. Cumplir las demás actividades relacionadas con el objeto del contrato que sean acordadas con el supervisor.

    - Entrege reports preliminares y diapositivas relacionados con los proyectos que se han desarrollado y culminado a la fecha para revisión del coordinador del equipo de Big Data.
    - Desarrolle el Producto explicando el detalle técnico y los resultados de los proyectos desarrollados durante la ejecución del contrato. (Análisis de redes para el CONPES animalista, identificación de la red de empresarios colombianos para la estrategia de comunicaciones, desarrollo de la herramienta que hace seguimiento en tiempo real de los precios agropecuarios reportados en las centrales mayoristas del país).

Mario Andres Morales Moreno
Cédula 79.722.799 de Bogotá D.C.

Contrato de Apoyo a la gestión

Apoyar a la Oficina de Informática del Departamento Nacional de Planeación – DNP, en la coordinación y generación de insumos, para realizar procesos de inteligencia de negocios, análisis de datos, el procesamiento y presentación de grandes volúmenes de datos (Big Data), involucradas en las iniciativas, pilotos y soluciones estratégicas de la entidad.

Mario Andrés MORALES MORENO

Informe N° 09 de 12
Contrato DNP-457-2017
30 de octubre de 2017
Dirección de Infraestructura y

DNP Departamento Nacional de Planeación

Contrato de prestación de servicios profesionales y de apoyo a la gestión, financiado con recursos ordinarios del Departamento Nacional de Planeación, con base en el artículo 81 del Decreto 1510 de 2013, bajo la supervisión de la Dirección de Telecomunicaciones y la Oficina de Informática del Departamento Nacional de Planeación.

30-10-2017

Recibí:
30-10-2017

En cumplimento con lo establecido en el contrato de prestación de servicios profesionales y de apoyo a la gestión No. 457 de 2017 celebrado entre el Departamento Nacional de Planeación (DNP) y Mario Andrés Morales Moreno, me permito entregar el producto de las siguientes actividades realizadas entre febrero 16 y 31 de octubre del presente año:

1. Apoyar a la Oficina de informática del DNP, en la coordinación, gestión de datos, ciencia de datos y en general el ciclo de vida de los datos de la entidad.

- Coordinación de los proyectos culminados a el 30 de octubre de 2017 (Ver Anexo 2)

Análisis exploratorio de datos abiertos en Colombia y en el contexto con respecto a otros repositorios de datos abiertos (GOVEDA)
Análisis de texto de los planes de desarrollo territorial
CONPES de protección de animales
Análisis a Registros Únicos
Análisis de proyectos de Regalías
AGRONET Análisis de los determinantes asociados a la generación de mayores ingresos en los hogares rurales de Colombia
Secretaria de Transparencia
Análisis de sentimiento de red de empresarios

2. Apoyar a la Oficina de Informática, en las iniciativas y soluciones desarrollados por las dependencias que impliquen la implementación de componentes de análisis de datos y que apoyen la toma de decisiones de nivel estratégico y operativo. (Ver Anexo 1)

Herramienta de Búsqueda de Datos Masivos (Big Data) para Mejora Normativa

3. Apoyar a la Oficina de Informática, en la elaboración de informes estadísticos sobre correlación de variables, analítica descriptiva y predictiva; de fuentes de información internas y/o externas de la entidad que apoyen la toma de decisiones estratégicas.

Bajo mi gestión proporcione soporte técnico y recomendaciones para el desarrollo e implementación de los tableros de control de sistema RIEPI y recomendaciones para implementación y finalización del sistema SUCOP.

4. Apoyar a la Oficina de Informática, en la elaboración de insumos matemáticos y estadísticos que soporten los informes estadísticos solicitados por la Oficina de Informática o por las dependencias de la entidad; utilizando las herramientas de software adecuadas.

- Uso de Bases de Datos MS SQL Server pára implementación del proyecto SIRH.
- Uso inicial y extensivo de MS PowerBI.
- Estudio y recomendación en la adquisición de OCR ABBYY para proyecto de reconocimiento de texto en el DNP.

- Uso de Ubuntu Linux en el grupo de Ciencias de Datos del DNP.
- Uso inicial y extensivo de Elasticsearch en prototipo de mejora normativa del DNP
- Recomendaciones de Seguridad informática desde el punto de vista de Ciancias de Datos.

5. Apoyar a la Oficina de Informática, en la consolidación de datos provenientes de fuentes internas y/o externas de la Entidad para facilitar las tareas de análisis.

- Adquisición y análisis de información inicial proveniente de WAZE para proyectos de movilidad internos.
- Recreación e implementación de algoritmos en Sistemas de Información geográficos (SIG) para el proyecto del ciclo del agua SIRH

6. Apoyar a la Oficina de Informática, en la gestión de información, resultante de procesos y procedimientos de captura, almacenamiento, procesamiento y análisis de datos.

- Adquisición de información inicial proveniente de WAZE para proyectos de movilidad internos.

7. Apoyar a la Oficina de Informática, en la administración, configuración, programación y operación de herramientas de apoyo para la gestión de información asociada análisis de datos.

- Gestión de permisos para la instalación de servidor de investigación del grupo de ciencia de datos.
- Gestión de permisos para la instalación de elasticsearch para el prototipo del proyecto de mejora normativa.

8. Apoyar a la Oficina de Informática, en las transferencias de conocimiento dirigidas a las áreas del DNP, relacionadas con las implementaciones realizadas y en los casos que se requiera.

- Dos presentaciones en el seminario técnico del DNP, presentación de resultados en la universidad externado de Colombia y presentación el open data day de Bogotá, Colombia.

9. Integrar Comités Asesores-Evaluadores asignados por el supervisor del contrato en relación con los procesos de contratación que requiera la Oficina de Informática del DNP.

- Web scraping de precios en páginas web de retailers (grandes superficies).
- Participación en reunión con CCP WAZE para colaboración de proyectos de analítica de datos.
- CAOBA (Proyecto de Agua)

- Proyectos de la Dirección de Seguimiento y Evaluación de Políticas Públicas (DSEPP)
- Bien Público Regional (BPR) del BID
- Visualización de datos Índice Departamental de Innovación para Colombia
- Convenio con el Massachusetts Institute of Technology (MIT) Predicción de supervivencia y muerte de empresas y agentes en sistema SISBEN.
- Fiscalía General de la Nación (FGN).
- Estudio de perfiles poblaciones, flujos turísticos y decisiones de consumo de servicios turísticos.
- Participación en reunión con IDEAM para colaboración de proyectos de analítica de datos en el SIRH.
- Asistencia a reunión de presentación de resultados sobre el proyecto de protección animal realizado en los meses de febrero y marzo ante la comisión de estadística e indicadores del Ministerio TIC.
- Participación en reunión con la Dirección de Desarrollo Social del DNP y para definición del proyecto Big Data y Machine Learning del DNP puede trabajar en colaboración con ellos en el manejo de datos provenientes de fuentes como ADRES (antes FOSYGA).
- Participación en reunión con la DDE y el grupo de comunicaciones del DNP para determinar la estrategia a desarrollar para la Gran Encuesta de empresarios de La Republica.
- Asistencia a reunión de presentación de resultados obtenidos durante la captura de información vía Twitter los días 29 y 30 de agosto denominada "red de empresarios en Colombia" ante el Director General Luis Fernando Mejía Álzate el día 22 de septiembre de 2017.
- Presentación de Resultados en el evento de BIG Data del DNP en la Universidad Externado de Colombia

10. Cumplir las demás actividades relacionadas con el objeto del contrato que sean acordadas con el supervisor.

- Entrega del presente informe de producto con proyectos culminados a la fecha para revisión de supervisores del contrato en STEL y la oficina de Informática.
- Continuación de los siguientes proyectos, con horizonte a 2018 y futuro:

Proyecto de Movilidad Bogotá y Principales Ciudades
Análisis de Texto Planes Nacionales Desarrollo
Análisis de Sentimiento, Percepción Empresarial
Ministerio de Salud, ADRES
SIRH Ciclo del Agua
Generación de reportes de información interna del DNP (ORFEO) para medir información que se está generando internamente.

ANEXO 1

1. Herramienta de Búsqueda de Datos Masivos (Big Data) para Mejora Normativa

Participantes: Mario Morales y José Librado Mejía. Bajo la coordinación de Cesar Pabón

1. Objetivos de la Política de Mejora Regulatoria

Busca mayor bienestar económico y social por medio de herramientas que permitan una producción normativa de calidad y coherente



2. Antecedentes, ¿Cuánto le cuesta la regulación a Colombia?

La excesiva regulación tiene un impacto en la capacidad del Estado para hacerla cumplir y en la competitividad del país

Foro Económico Mundial 2016 - 2017	Índice de competitividad	61 de 138	"En materia institucional, Colombia pasó de la posición 79 a la 112 en la última década, quedando al nivel de países como Nigeria o Pakistán. Factores como la corrupción, la falta de independencia judicial, la carga regulatoria, el crimen organizado y el terrorismo, explican este resultado."
	Índice de carga regulatoria	124 de 138	
Rule of Law 2016 - 2017	Índice Rule of Law	71 de 113	Colombia ocupa el puesto 15 de 30 en la región en el Índice de Cumplimiento Regulatorio que mide la capacidad de los países en hacer cumplir la regulación
	Índice cumplimiento regulatorio	51 de 113	

Fuente: Consejo Privado de Competitividad 2016

3. Implementación del Proyecto en el DNP

Recolección de Información
Se incorporo información disponible de los Diarios Oficiales de la Imprenta Nacional

→

Identificación de patrones
A través de un proceso de Data-mining se permite explorar la base de datos con el objetivo de encontrar patrones repetitivos, tendencias o reglas

→

Estandarización
Se asoció la información de los diarios con las instituciones del orden nacional disponibles en el SIGEP

→

Base de datos
Permite realizar diferentes categorizaciones de la información para la Rama Ejecutiva del orden Nacional

Resultado

OFRECE:

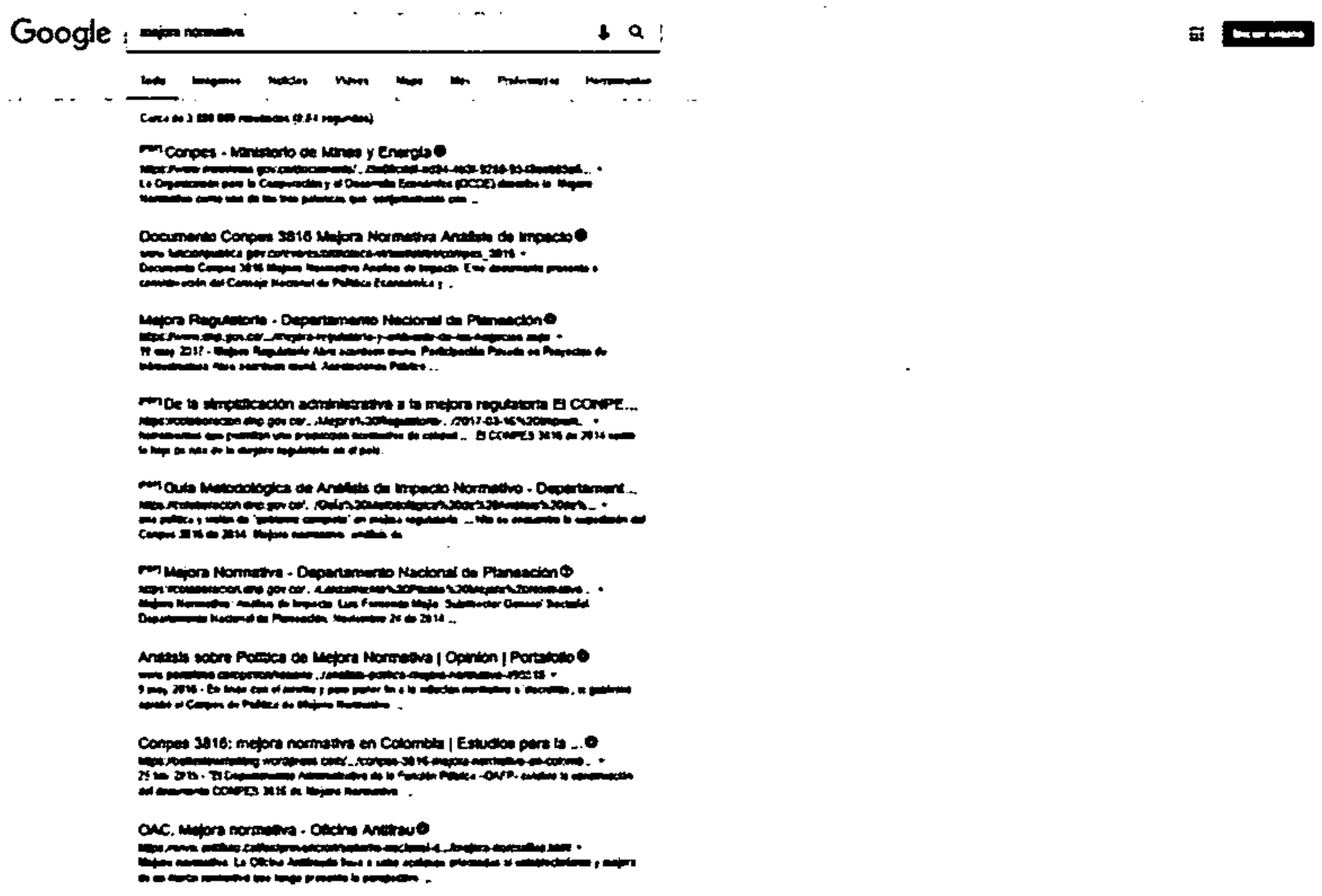
- Información disponible desde el 2000
- Datos con diferentes frecuencias: anuales, mensuales, semanales e incluso diarias
- Categorizarlos
- Restringir los datos geográficamente
- Actualización tiempo real

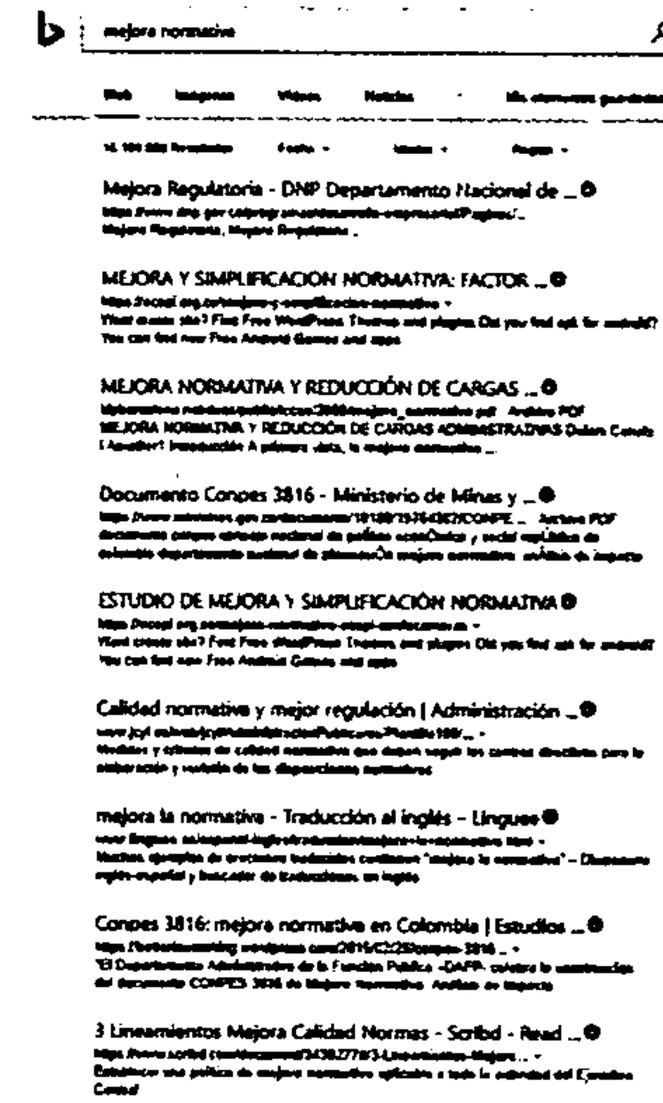
Base de datos
Rama Ejecutivo Nacional (2000-2016)
94.748 observaciones (actos administrativos por instituciones)
4 clasificaciones de la información

25 sectores administrativos	Clasificación - Sustancial No Sustancial	21 tipos de normatividad emitida	23 clases de naturaleza jurídica para instituciones

4. Motores de Búsqueda

Herramientas como Google y Bing permiten consultar grandes cantidades de información en tiempos óptimos




mejora normativa

Mejora Regulatoria - DNP Departamento Nacional de _

MEJORA Y SIMPLIFICACION NORMATIVA: FACTOR ...

MEJORA NORMATIVA Y REDUCCIÓN DE CARGAS ...

Documento Conpes 3816 - Ministerio de Minas y ...

ESTUDIO DE MEJORA Y SIMPLIFICACIÓN NORMATIVA

Calidad normativa y mejor regulación | Administración ...

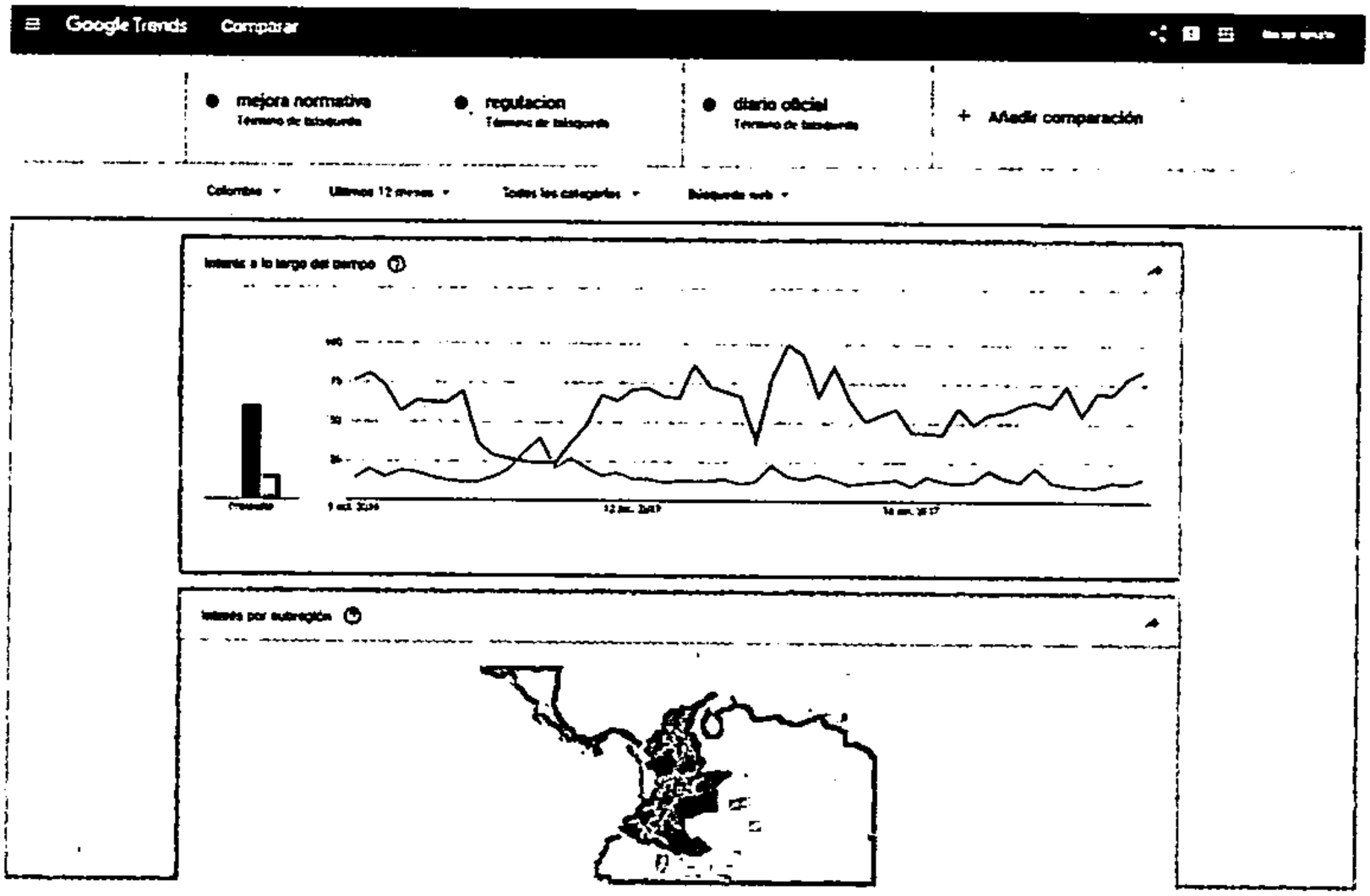
mejora la normativa - Traducción al inglés - Linguee

Conpes 3816: mejora normativa en Colombia | Estudios ...

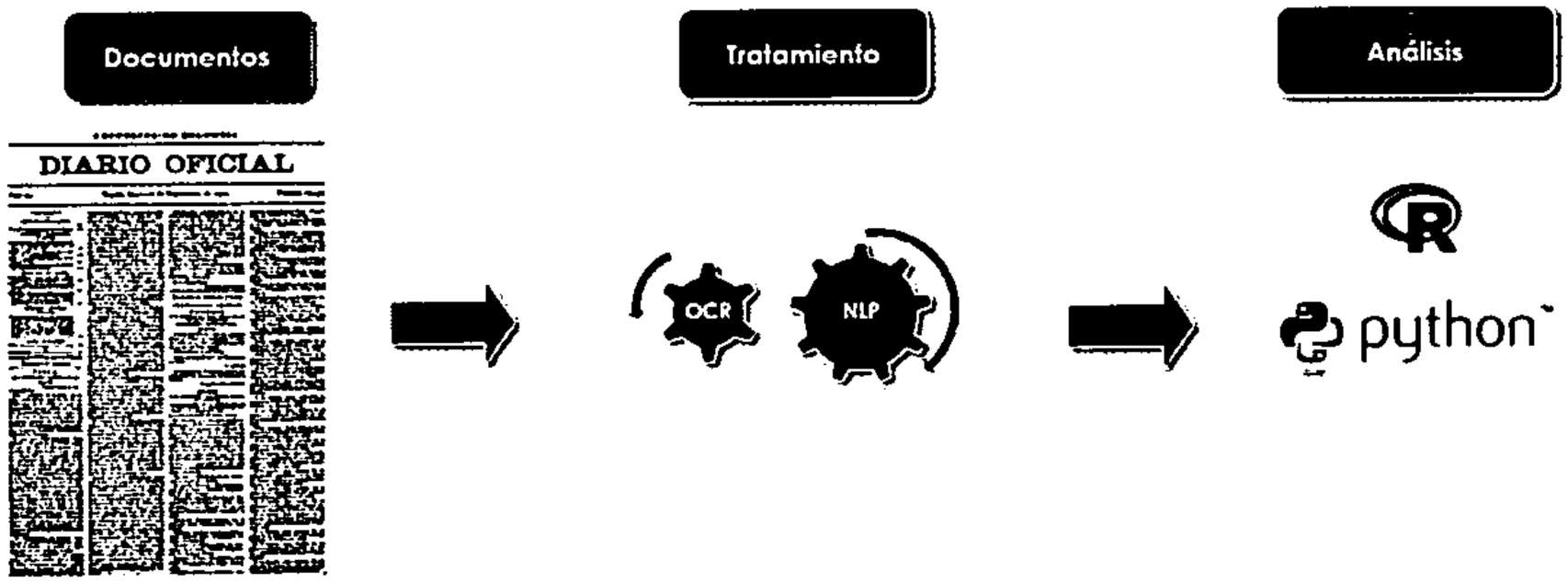
3 Lineamientos Mejora Calidad Normas - Scribd - Read ...

5. Tableros de Control y Análisis

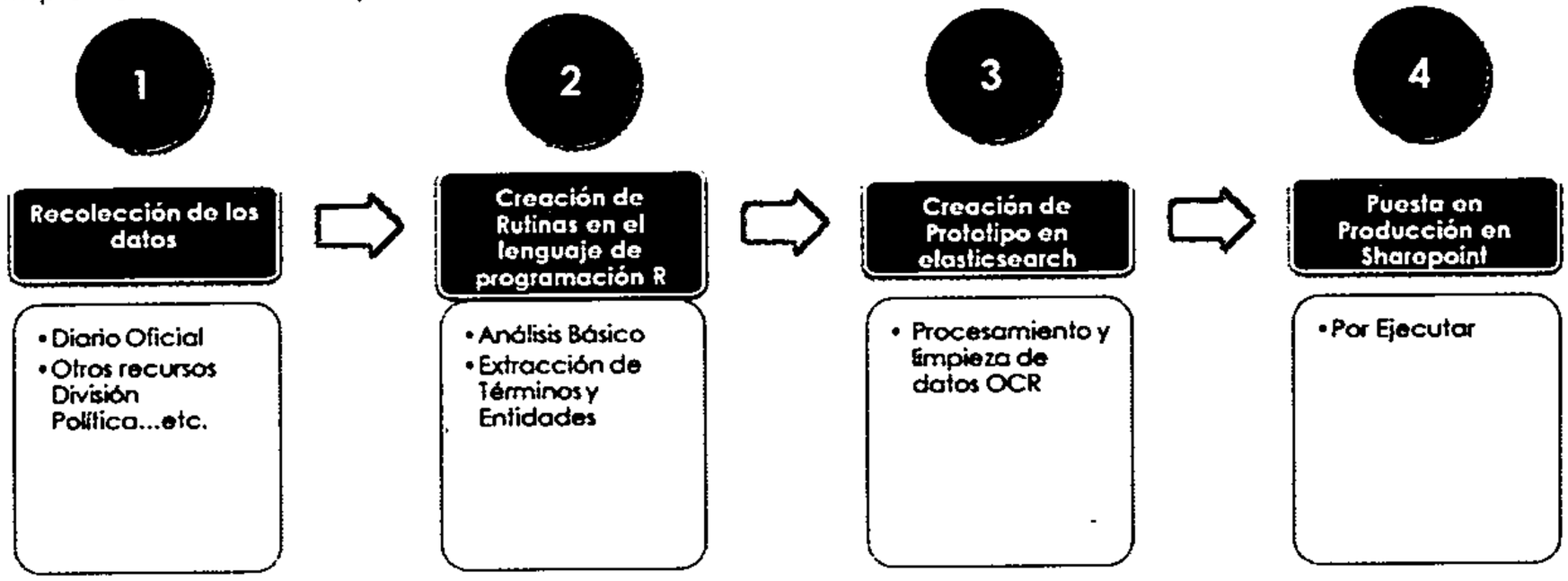
Por ejemplo: Google Trends permite sintetizar información acerca de palabras clave a diferentes frecuencias temporales

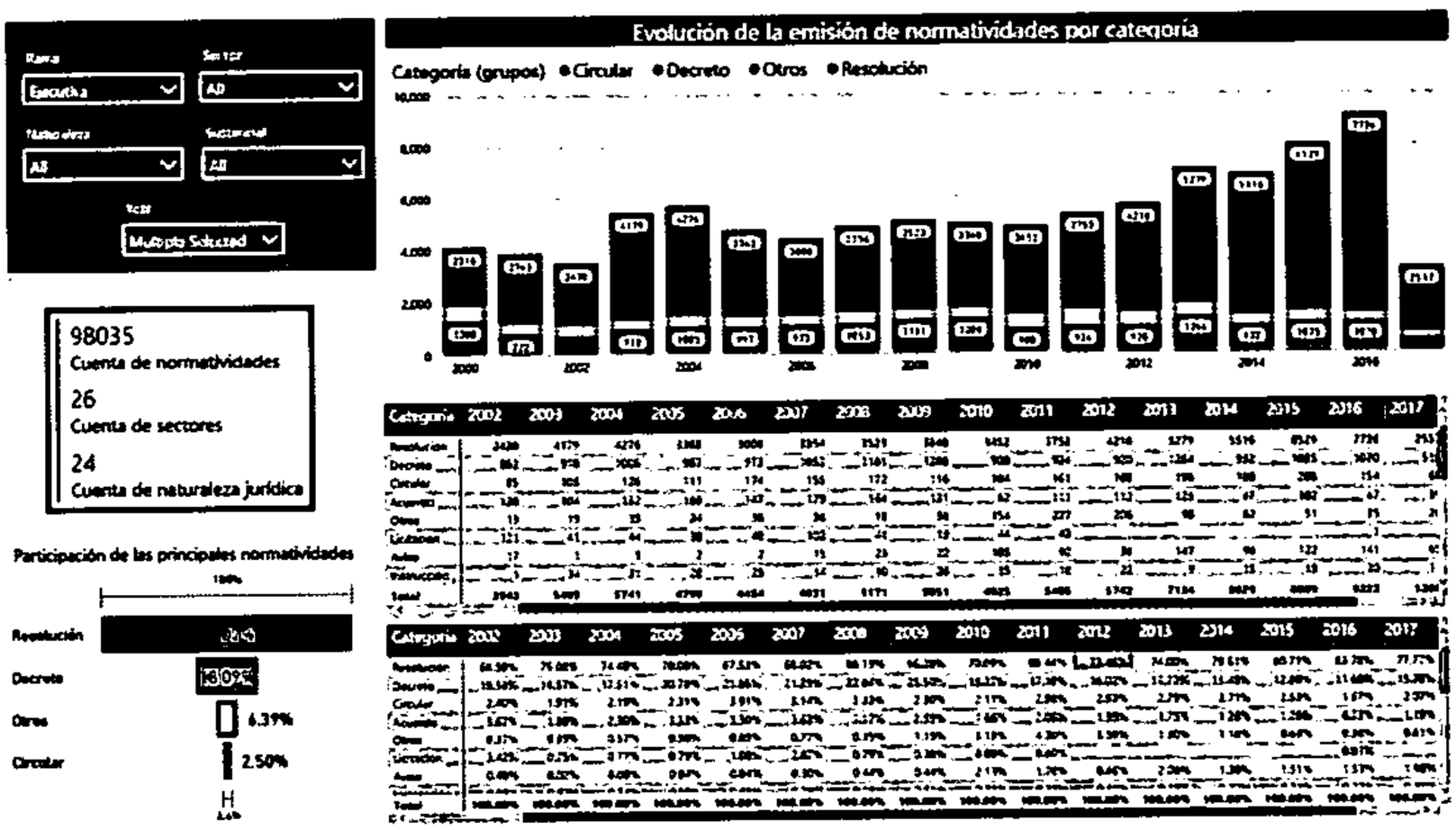


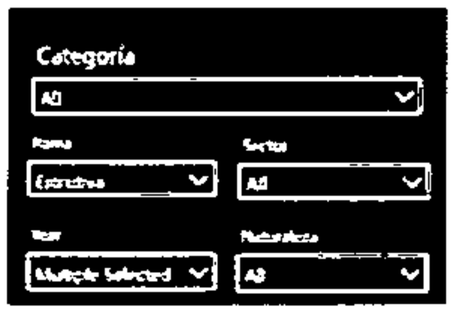
6. Serie de pasos para la Implementación del motor de búsqueda:

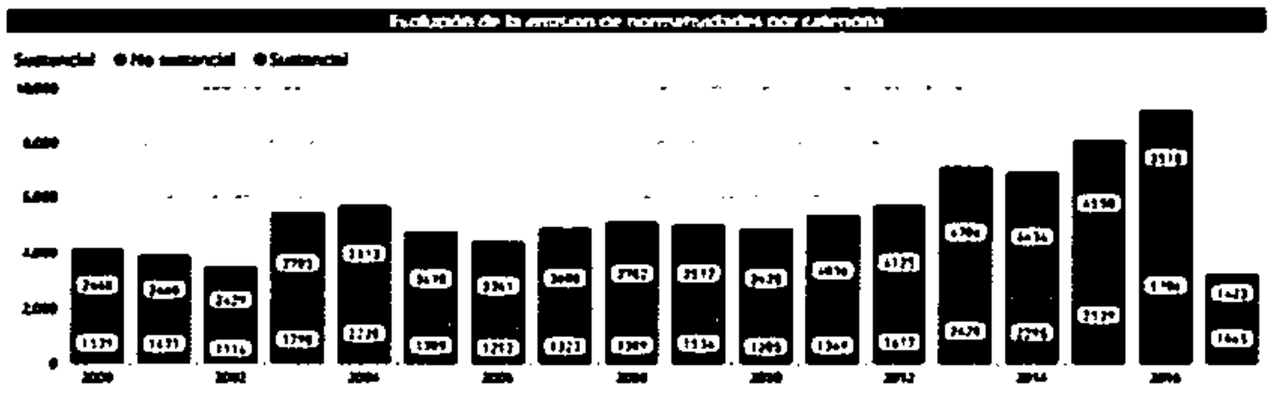


7. Implementación del Proyecto





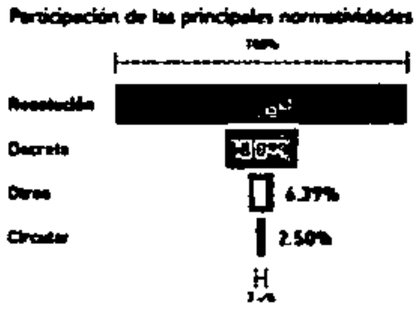


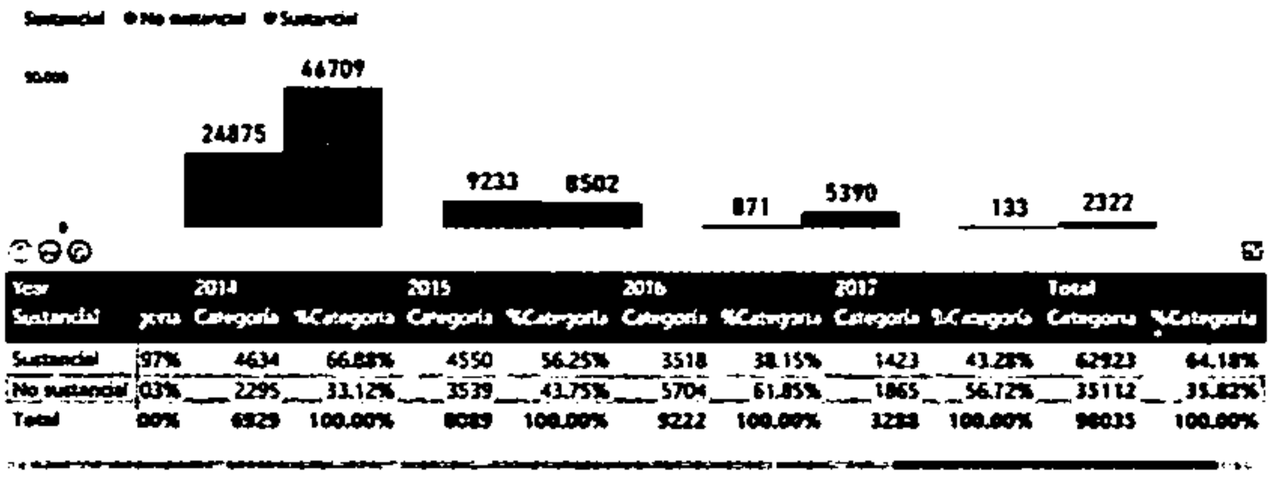


Participación por categoría sustancial

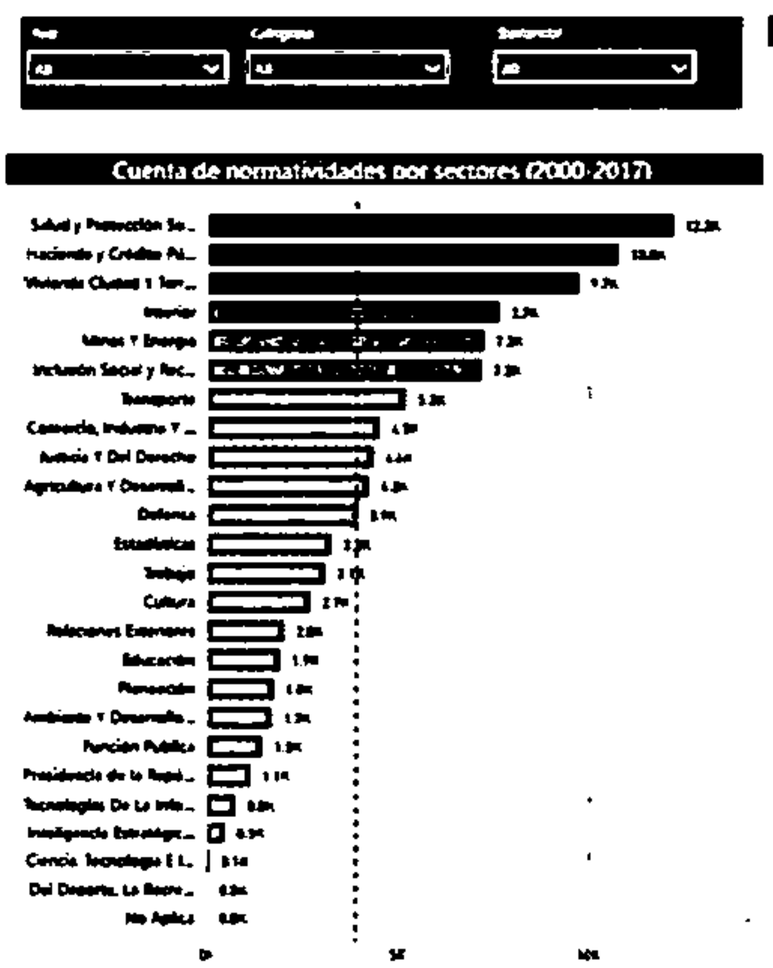


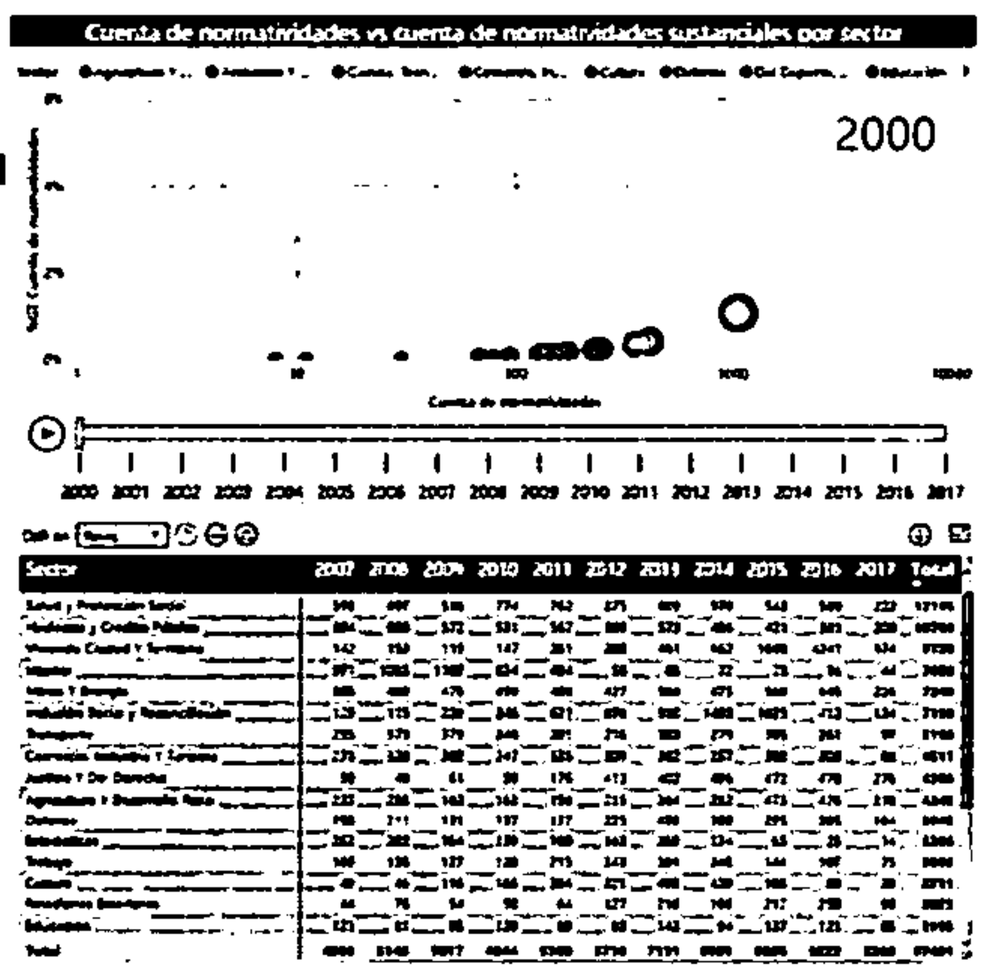
Participación de las principales normatividades





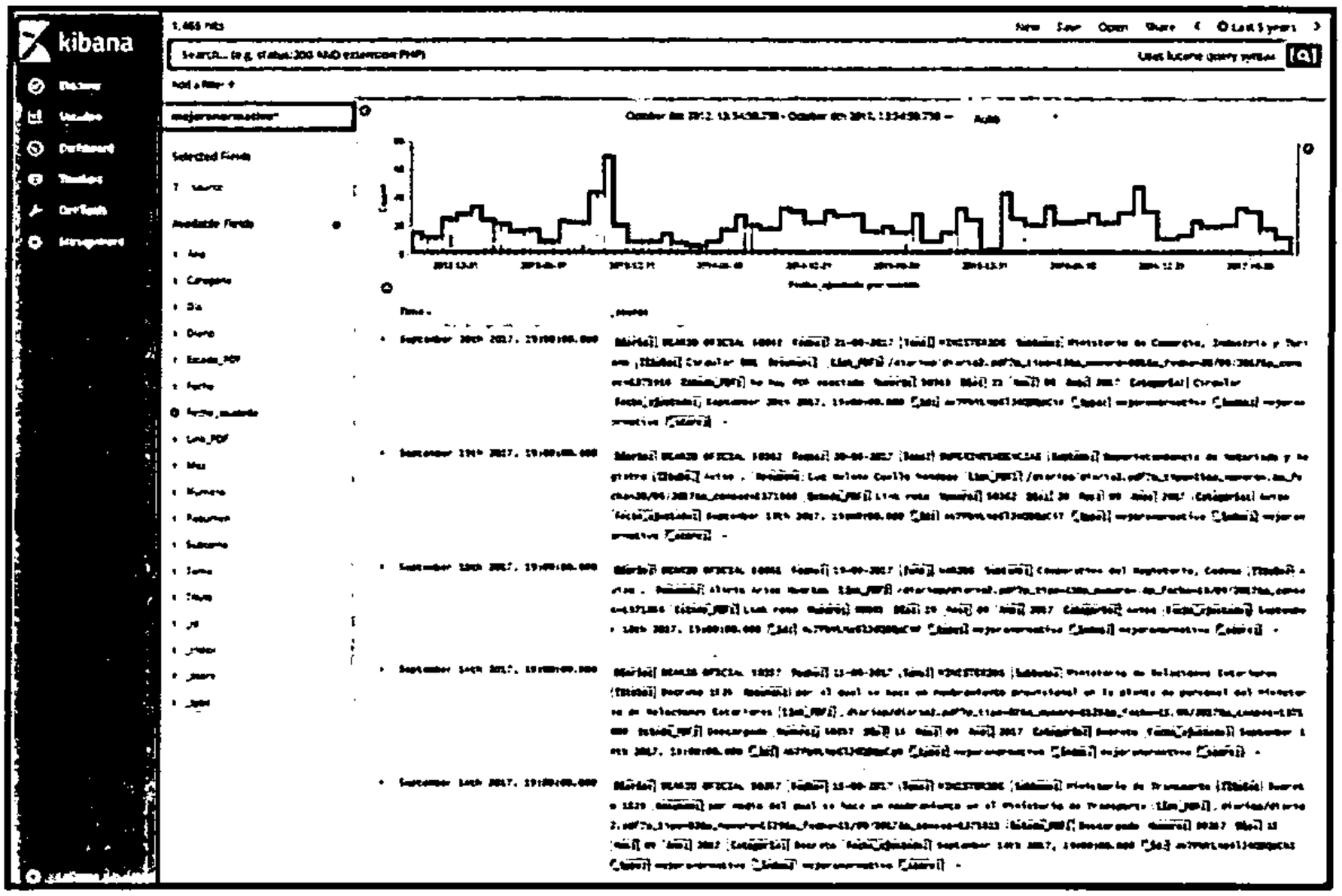
Year	2014		2015		2016		2017		Total	
Sustancial	%Categoría	Categoría	%Categoría	Categoría	%Categoría	Categoría	%Categoría	Categoría	%Categoría	Categoría
Sustancial	97%	4634	66.88%	4550	56.25%	3518	38.15%	1423	43.28%	62923	64.18%
No sustancial	03%	2295	33.12%	3539	43.75%	5704	61.85%	1865	56.72%	35112	35.82%
Total	00%	6929	100.00%	8089	100.00%	9222	100.00%	3288	100.00%	98035	100.00%





Sector	2007	2008	2009	2010	2011	2012	2013	2014	2015	2016	2017	Total
Salud y Protección Social	590	697	586	774	762	675	609	599	543	589	222	12196
Hacienda y Crédito Público	884	888	572	531	567	888	529	486	423	381	208	10700
Vivienda Ciudad Y Territorio	142	159	119	147	261	288	491	462	1648	4347	534	9728
Interior	891	1088	1188	654	484	55	65	22	25	94	44	5000
Minas Y Energía	885	888	478	659	486	427	588	471	568	648	226	7249
Inclusión Social y Reconciliación	529	515	229	346	671	899	985	1488	1825	713	534	7198
Transporte	255	571	379	848	381	716	888	579	595	261	97	5150
Comercio, Industria Y Turismo	279	538	388	347	535	889	382	257	388	308	96	4511
Justicia Y Del Derecho	98	48	61	98	176	413	427	486	472	478	276	4388
Agricultura Y Desarrollo Rural	233	288	163	163	158	235	384	282	473	476	278	4348
Defensa	988	711	191	197	177	225	488	388	295	385	164	4048
Estadísticas	282	288	184	139	988	163	288	334	65	28	34	4386
Trabajo	188	186	127	128	215	143	384	348	144	187	75	3888
Cultura	48	46	116	166	284	471	488	438	388	88	28	3751
Relaciones Exteriores	66	76	58	98	66	127	218	188	217	238	98	3883
Educación	121	65	88	139	88	88	142	94	137	121	88	3998
Total	4888	5148	5817	4844	5388	5778	7191	6888	6888	6822	3388	97484

8. Prototipo con elasticsearch y kibana



9. Conclusiones
    - Se determino la necesidad y el uso de tratar información no estructurada por medio del uso de un motor de búsqueda.
    - Esta herramienta permitirá producir nueva regulación adecuada, eficaz y transparente y de la misma manera habilitará a el gobierno y sus instituciones la revisión de la normatividad vigente.
    - La construcción de tableros de control permitirá el descubrimiento de anomalías y permitirá la construcción de modelos predictivos basados en el contexto del contenido regulatorio producido.
    - Se implementará en Producción en el futuro cercano 2018-2019.

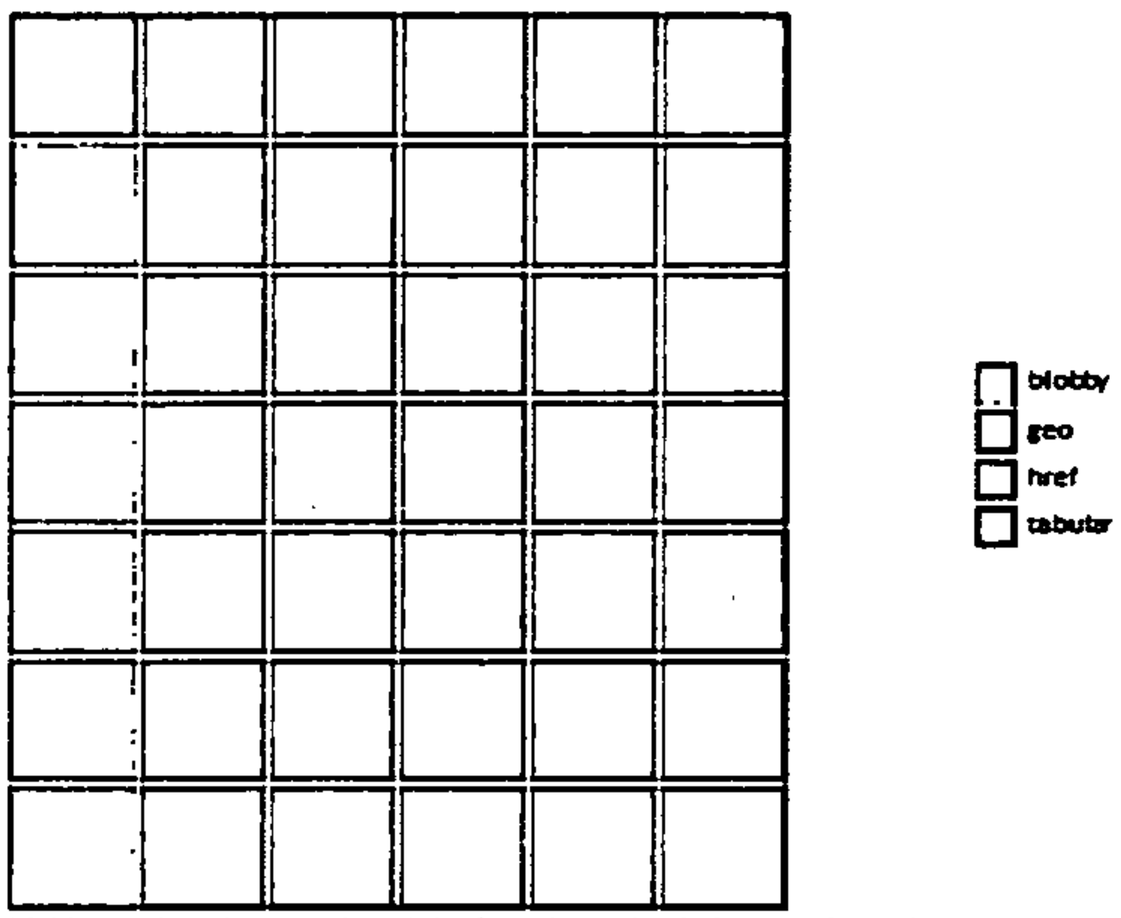
ANEXO 2

2. Government Data Analysis (GOVEDA)

Participantes: José Zea y Mario Morales
Objetivo: facilitar el consumo de datos ubicados en portales open data
Actividades:

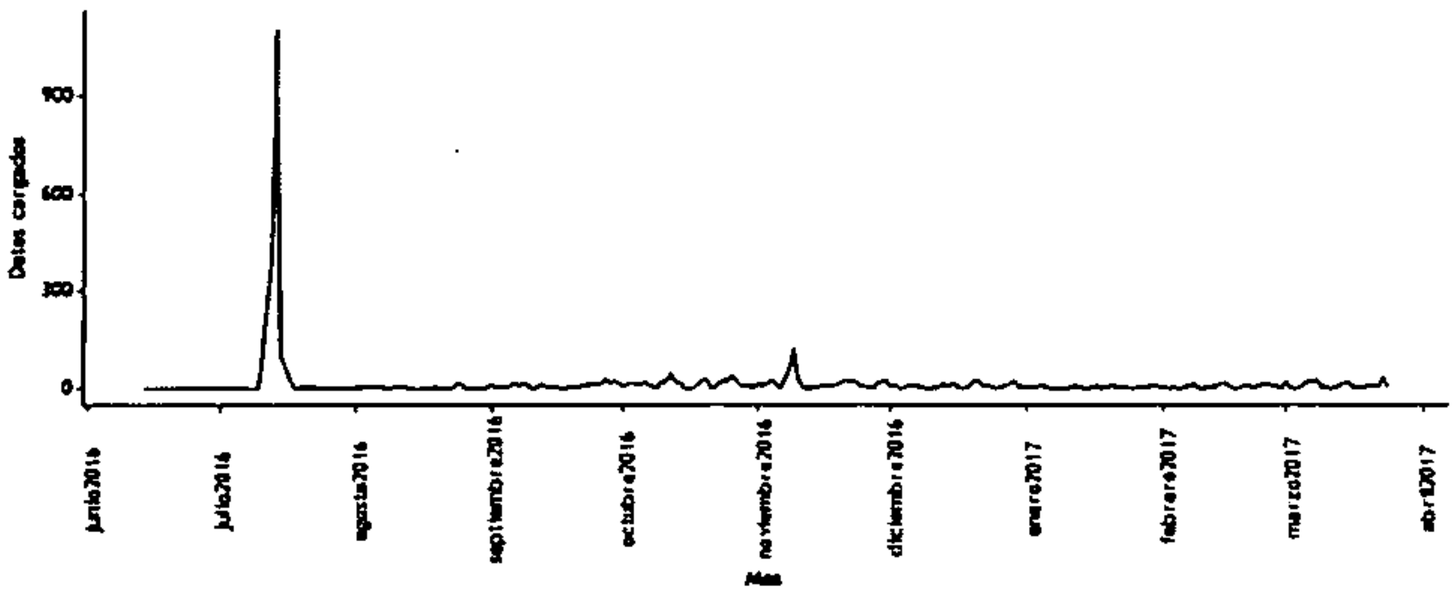
- Desarrollo de scripts para leer datos de portales de datos abiertos alojados en los portales de datos abiertos *Socrata* y CKAN.
- Facilitar la búsqueda de los datos disponibles en los portales por palabras claves o etiquetas.
- Generar reportes básicos con los metadatos disponibles de los portales de datos abiertos.
- Conocer el número de ficheros disponibles y algunas estadísticas generales que den cuenta de la información disponible en los portales de datos abiertos.

Para ilustrar la utilización del portal de datos abiertos se muestran algunos de los análisis que se desarrollan con GOVEDA:
La mayoría de los datos en el portal de datos abiertos de Colombia son de tipo tabular:

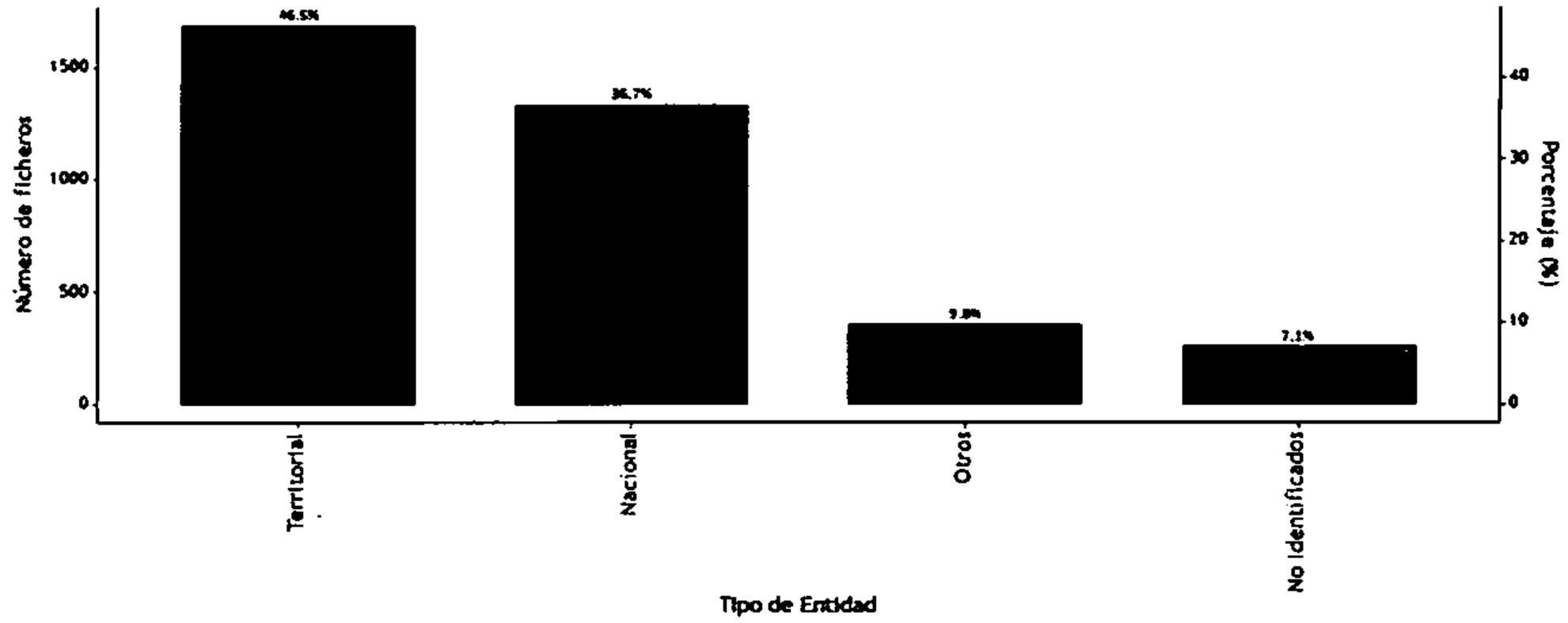


blobby: 174(3.9%); geo: 57(1.3%); href: 577(12.8%); tabular: 3697(82.1%)

La generación de datos abiertos tuvo un pico en Junio de 2016 cuando se creó la iniciativa del portal de datos abiertos:



Las entidades que más producen información son las del orden territorial con cerca de la mitad de los datos publicados (46,6 %), le siguen las entidades del orden nacional con el 36,7 de datos publicados.



Las funciones generadas para analizar los portales de datos abiertos son:

get_metada

description: a detailed metadata of open datasets (extract json and basic list of datasets available from a Socrata domain

search_data: Shows a list object with available datasets by keywords/tags. The list contains four things:

eda_opendata: Generate a basic report in rmarkdown or shiny with selected datasets from open data portal.

sizeTabularData: Compute number of rows and columns, size and elapsed system time of every selected dataset

ReadTabularData

Description read a tabular dataset of open portal in R. . Fetch a detailed metada of open datasets (extract json and basic list of datasets available from a Socrata/CKAN domain)

Usage get_socrata_metada(url)

Arguments: url - A Socrata/CKAN URL. This simplest points to the site root.

Details: A R data frame containing a listing of datasets along with detailed metadata. Next field are preserved for every open dataset:

id, name, attribution, averageRating, category, createdAt, description, displayType, downloadCount, hideFromCatalog, hideFromDataJson

viewCount, viewLastModified, viewType, columns, grants, license, metadata ratings disabledFeatureFlags, accessLevel, landingPage, issued, @type, modified

keyword, contactPoint, publisher, identifier, title, distribution, theme

search_data

Description Shows a list object with available datasets by keywords/tags. The list contains four data frames:

- Tabular datasets: A data frame with four columns (dataset name, attribution (public agency), number of rows and columns size of dataset (Mb) and elapsed time for R export.

- Geo: A data frame with general information about geo file.

- Href: A data frame with general information about href file.

- Blobby: A data frame with general information about blobby href file.

Usage

search_data(metadata, tags)

Arguments

metadata: A data frame get from get_metadata function

tags: A character vector with keywords and/or tags required for searching in dataset name, tag field or public agency field

3. Análisis Planes de Desarrollo Territorial

Participantes: José Zea, Daniel Rojas y Mario Morales
Objetivo: encontrar planes de desarrollo territorial similares y construir grupos temáticos de los mismos.
Actividades:

1. Lectura de la información contenida en 106 Planes de Desarrollo Territorial
2. Depuración de la información (lematización, e identificación de las palabras más frecuentes).
3. Desarrollo de un análisis de k-medias y de caracterización con la matriz de palabras.
4. Cálculo de medidas de similitud con la matriz de palabras.
5. Similitud de palabras con palabras claves preestablecidas.

El análisis de los planes de desarrollo territoriales (PDT) se realizó por medio del procesamiento de Lenguaje Natural, entendido como aquella forma mediante la cual los computadores pueden analizar textos en leguaje convencional (español, inglés, portugués, entre otros) y no artificial (lenguaje de programación o alfanumérico) (Bird, Ewan y Edward 2009), a través del lenguaje de programación en *Python*[1].
Por medio de esta herramienta se realizó el pre-procesamiento del texto, que incluyó: convertir todo el texto a minúsculas y leer solo los caracteres alfabéticos, eliminando así números, símbolos, viñetas y signos de puntuación; además, se eliminaron aquellas palabras recurrentes en el uso del lenguaje pero que solo dan estructura y lógica al mismo como conjunciones, artículos y preposiciones, conocidos en lenguaje de programación como *stopwords.* El listado de *stopwords* en español que tienen las librerías de *Python* es muy incipiente y por esa razón se construyó un inventario de palabras que igualmente fueron excluidas de la lectura de los textos, y que en el código fueron llamadas *filtros (*los filtros incluyeron adjetivos numerales, demostrativos e indefinidos y números
Después de ese ejercicio de limpieza de textos se construyó, con las diferentes áreas técnicas del DNP, un listado de temáticas (sectores básicos, grupos poblacionales, sectores económicos) que son relevantes al momento de analizar el contenido de los PDT y que agrupan distintas palabras asociadas a la temática (para ver las temáticas construidas y sus palabras asociadas ver Anexo No.1); por ejemplo, el sector de educación está compuesta por palabras como: docentes, aulas, estudiantes, colegios, maestros, pruebas SABER, Ministerio de Educación, educación básica, educación media, educación superior, entre otros. Como alguna de estas nociones están constituidas por 2 o más palabras consecutivas, hubo la necesidad de construir un listado de palabras compuestas, lo que en lenguaje de programación se llama *n-gramas* y en el código fue denominado como *grupos* para que en el procesamiento se analizaran conjuntamente. Una vez se realizado el procesamiento del texto, se hizo un conteo de frecuencias de las palabras que componen el PDT, es importante resaltar que para el análisis de frecuencias los *n-gramas* son tomados como una sola palabra.

---

[1] Cabe anotar que los siguientes PDT fueron procesados en R debido a que no fue posible hacer el preprocesamiento del texto en Python: Duitama, Boyacá; Caldas, Boyacá; San Antonio del Tequendama, Cundinamarca.

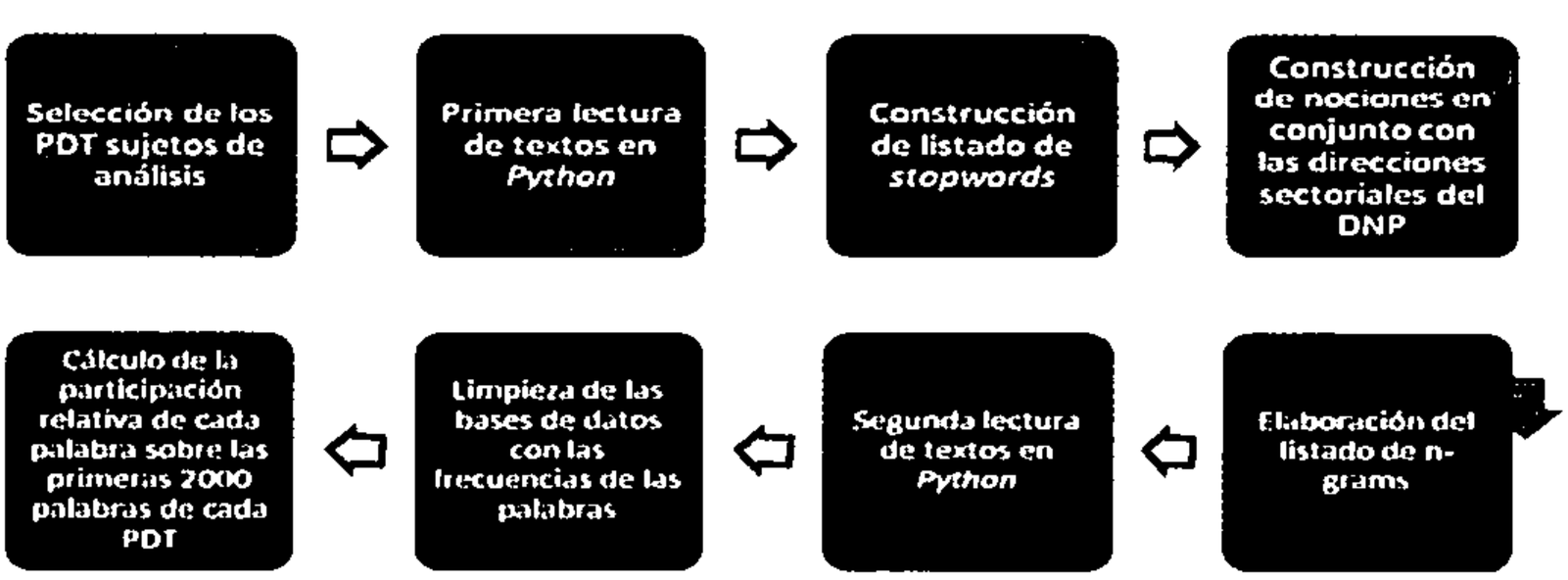
Terminado el procesamiento del texto y el conteo, el programa arroja como resultado un listado con todas las palabras que componen el PDT y su frecuencia. A este *output* se eliminan las palabras o se suma la frecuencia del término mal escrito al correcto. Con el listado de palabras y frecuencias depurado, y con el fin de hacer un indicador comprable para todos los municipios y departamentos del país, independientemente de la extensión del PDT, se estableció que el análisis se iba a realizar sobre las 2000 palabras más frecuentes, puesto que 2000 palabras representan cerca del 85%[2] del total de palabras de los PDT y, en una buena porción de los planes, estas 2000 tienen frecuencias mayores a 1.

El análisis de las frecuencias se obtiene de la participación relativa de cada término sobre las primeras 2000 palabras, para calcular la participación se divide la frecuencia de cada término sobre la sumatoria de las frecuencias de las 2000 palabras más repetidas y se multiplica por 1000 (frecuencia/∑frecuencia 2000) *1000. Para términos de comparación, las palabras más frecuentes de los planes tienen, en promedio, una participación de 18.

Teniendo las participaciones relativas, lo que se hizo fue construir una base de datos con las temáticas y nociones por temática, con el fin de identificar la importancia de los ejes estratégicos del Plan Nacional de Desarrollo (PND), los sectores básicos, el enfoque poblacional y otros temas relevantes para el desarrollo del país en los PDT. Los resultados se presentan por el total de las temáticas, que no es más que la sumatoria de las participaciones relativas de los conceptos que la integran. Al final se evaluaron 20 temáticas y 5 grupos poblacionales.

Asimismo, a partir del listado de las 2000 palabras más frecuentes se grafica la nube de palabras o *WordCloud* para cada uno de los municipios y departamentos. En el *WordCloud* el tamaño de la palabra refleja la frecuencia con la que se repite la palabra dentro del PDT, así las palabras de mayor tamaño son las que más se repiten en el texto.

Gráfico 1.
Metodología utilizada para este análisis



Por otra parte, y con el fin de identificar si había diferencias significativas en la importancia que daban los departamentos y municipios a las temáticas y poder agrupar a los entes territoriales según la importancia con la que trataban los temas en su PDT,

[2] Natagaima, Tolima y Mercaderes, Cauca tienen menos de 2000 palabras.

además de tratar de entender a qué se debían esas diferencias, se aplicó un método de agrupamiento llamado *k-means*. Este método usa iteraciones de agrupamiento aleatorios hasta estabilizarse y lograr uno que minimice la inter-varianza dentro de los elementos de un grupo (*cluster)* y maximice la intra-varianza entre los diferentes *clusters*, asegurando así que los grupos sean lo más homogéneos en su interior y lo más disimiles entre ellos. Este método de agrupación utiliza una prueba de diferencias de medias (entre el de cada término del PDT con la de cada término del total de PDT evaluados) para conformar los grupos.

El insumo sobre el cual se realiza la clasificación es la matriz de *matriz de palabras – documentos.* La matriz se encuentra conformada en las filas por las palabras que satisfacen los siguientes criterios, i) palabras resultantes de la limpieza explicada anteriormente (eliminación de *stopwords*, puntuaciones, etc.), ii) palabras cuya ocurrencia en los PDT es mayor a 2000 y palabras, iii) Palabras cuya aparición es significativa en todos los documentos (en total se retiene para el análisis 5582) . En las columnas se encuentran cada uno de los PDT (106 analizados), los valores de las celdas ($n_{i,j}$) corresponden a la frecuencia de ocurrencia de cada palabra en el documento respectivo.

Palabras/Doc	D_1	D_2	⋯	D_{106}
P_1	$n_{1,1}$	$n_{1,2}$	⋯	$n_{1,106}$
P_2	$n_{2,1}$	$n_{2,2}$	⋯	$n_{2,106}$
⋮	⋮	⋮	⋱	⋮
P_{5582}	$n_{5582,1}$	$n_{5582,2}$	⋯	$n_{5582,106}$

Las frecuencias son relativizadas con respecto a cada uno de los documentos (en documento se obtiene la frecuencia relativa de cada una de las palabras), la suma de las frecuencias por columna es 1.

Palabras/Doc	D_1	D_2	⋯	D_{106}
P_1	$f_{1,1}$	$f_{1,2}$	⋯	$f_{1,106}$
P_2	$f_{2,1}$	$f_{2,2}$	⋯	$f_{2,106}$
⋮	⋮	⋮	⋱	⋮
P_{5582}	$f_{5582,1}$	$f_{5582,2}$	⋯	$f_{5582,106}$
Total	1	1	⋯	1

Posteriormente con la matriz obtenida se lleva a cabo un algoritmo de clasificación no supervisado de k-medias el cual agrupa las palabras de acuerdo a su frecuencia de aparición relativa en los documentos. Las clases construidas se describen de acuerdo a la frecuencia en que ocurren en cada uno los documentos. Por ejemplo, la clase 21 está conformada por las siguientes:

Palabra	Frecuencia	Grupo
Energética	182	21
Actuaciones	171	21
Alcaldías	168	21
Pedagogía	167	21
Aula	166	21
Tolerancia	165	21
Bicicleta	125	21
Cerros	116	21
Crecer	61	21
Maternidad	60	21
Extranjeros	59	21
.	.	.
.	.	.
.	.	.

La aparición de las palabras de la clase 21 es estadísticamente más alta en la clase 21 que el corpus de los PDT. Las palabras de la clase 21 tiene una frecuencia de aparición de 39 veces por cada 10000 palabras en el documento, las mismas palabras tienen una aparición promedio de cerca de 18 veces en todos los PDT, por lo tanto, las palabras que aparecen en esta clase tienen una propensión de utilización de más del doble en Bogotá que en los PDT en general. Los dos promedios (media del grupo y media global) son comparados mediante la prueba estadística *t- student* para evaluar si existen diferencias significativas en los dos promedios.

PDT	Grupo	Valor t	Media grupo	Media global	Nro. Palabras
Bogotá	21	2,117	0,039	0,0179	55

Finalmente, las clasificaciones obtenidas se contrastan con las temáticas construidas.
Para poder llevar a cabo estas agrupaciones hubo primero que realizar un proceso llamado lematización a la base de datos de las frecuencias. La lematización es un proceso lingüístico que consiste en regresar las palabras de su formal flexionada (plural, conjugada, en femenino) a su lema, algo así como llevar las palabras a su raíz, a la manera como se encuentra en un diccionario. Así, por ejemplo, estudiando pasa a estudiar; educativos, educativas y educativa pasa a educativo; y en nuestro caso, a las palabras que tenían problemas de ortografía se lematizó a la palabra correctamente escrita, como en el caso de las palabras sin tilde como educacion o participacion. Esto con el fin de que la agrupación por temáticas sea más ajustada. Una vez se identifica a qué se lematiza cada palabra, se suman en el lema las frecuencias de todas las formas flexionadas y se aplica el proceso de clustering.

De esta manera se identificaron X *clusters* al analizar los 128 PDT, y uno de los hallazgos más importantes de este análisis es que no existen diferencias significativas en la importancia relativa que dan los entes territoriales a temas como educación, salud, paz, conflicto, agua y saneamiento. Estos son temas que fueron incluidos en todos los PDT y debido a la ausencia de diferencias significativas en la importancia relativa, se podría decir que la inclusión de estos temas no se debe a factores "objetivos", tales como cobertura, resultados en pruebas, incidencia del conflicto, acceso, entro otras. Por el contrario, hubo 29 ET que si incluyeron algunos temas de manera diferenciada (ver Tabla 1), por ejemplo, Santa Marta trata el tema de turismo, y algunos municipios de la Guajira el cuidado prenatal y a madres gestantes.

Tabla 1. Temáticas diferenciadoras de los PDT

	Relación Estado-Ciudadano	Desarrollo Territorial	Desarrollo Económico	TICs	Desarrollo Rural	Reconciliación	Diversidad cultural	Minería	Hábitos saludables	Turismo
Antioquia	X	X	X							
Atlántico	X	X								
Boyacá	X	X	X							
Caldas	X		X	X						
Cauca							X		X	
Cundinamarca	X	X	X							
Guajira	X	X	X							
Magdalena			X							
Nariño					X	X				
Quindío			X							
Risaralda	X			X						
Santander	X		X							
Sucre	X	X								
Tunja	X		X							
Valle del cauca			X							
Riosucio						X				
Puerto Rico						X				
Río Iró					X					
Mapiripán					X					
Bojayá						X				
El Charco							X			
Mocoa	X									
El tambo								X		
El Carmen									X	
Mutatá									X	
Dibulla							X		X	
Uribia							X		X	
Manaure									X	
Santa Marta										X

Indicador de similitud de los PDT con el Plan Nacional de Desarrollo

Para cuantificar la similitud entre los PDT y con el Plan Nacional de Desarrollo con respecto a los términos utilizados se utilizar el coseno entre los vectores asociados a los PDT y el PND. El coseno genera una medida entre cero y uno, en donde 1 significa que los documentos son idénticos y 0 que los documentos difieren de forma máxima en cuanto a los términos utilizados.

Para esto se parte de la utilización de la matriz *documento-término*, la cual contiene la frecuencia de los términos utilizados en cada uno de los PDT (una vez realizada el preprocesamiento del texto: limpieza de *stopwords*, eliminación de puntuaciones, etc.). En las filas se referencia los PDT y en las columnas cada uno de los términos utilizados el Corpus.

Se calcula como medida de similaridad el coseno conformado por cada una de los vectores fila (documentos), esta medida tiene como efecto eliminar el efecto del tamaño de texto de cada uno de los documentos. La similaridad entre el documento i el documento j está dada por el cociente entre el producto punto de la fila i-ésima y j-esima de la matriz de documento-frecuencia sobre sus respectivas normas

$$\textbf{Similaridad}(D_1, D_2) = \frac{D_1 \cdot D_2}{\|D_1\| \|D_2\|}$$

Se calcula la medida de similitud de cada PDT (fila de la matriz documento-término) con el vector de frecuencias de términos del PND. Se muestran los 5 PDT más similares y menos similares (en azul y rojo respectivamente) con el PND en cuanto a la frecuencia de los términos utilizados:

PDT	Índice
Cundinamarca	0,74
Manizales	0,69
Guajira	0,68
Caldas	0,68
Antioquia	0,66
Carmen_del_Darien (Chocó)	0,34
Santa Barbara Iscuande (Nariño)	0,33
Puerto Guzmán (Putumayo)	0,31
Puerto Leguizamo (Putumayo)	0,31
Córdoba (Bolivar)	0,30

4. Análisis De Redes Sociales Para El Conpes Animalista

Participantes: Daniel Rojas y Mario Morales

Introducción

La creciente pertinencia que ha venido adoptando el tema de protección animal a nivel internacional, no ha sido ajena a Colombia. En efecto, el presente Plan Nacional de Desarrollo (PND) "Todos por un nuevo país 2014-2018" incluye, por primera vez en un PND, esta temática (Contagio Radio, 2014). Con el fin de promover el respeto, cuidado e integridad de animales domésticos, el Consejo Nacional de Política Económica y Social (CONPES) consideró necesario desarrollar un documento otorgando lineamientos de políticas públicas que permitan lograr tal objetivo.

Además, se reconoció la vía digital como la más eficiente para realizar este documento. Teniendo en cuenta lo anterior, a continuación, se justifica la metodología que se llevará a cabo para recoger los insumos3 que hagan posible la elaboración de un documento CONPES sobre protección animal. Adicionalmente, se presentarán algunas conclusiones preliminares obtenidas de la investigación que se ha realizado hasta el momento.

Justificación

La migración hacia sociedades digitales y de la información es actualmente una realidad. Cada día el número de personas utilizando internet y apropiándose de las Tecnologías de la Información y la Comunicación (TIC) es significativamente mayor a nivel global (ITU, 2016). Esto hace que simultáneamente la cantidad de información que se produce y almacena a diario aumente a una tasa igualmente acelerada.

En particular, una poderosa fuente digital de información sobre los individuos son las redes sociales. Si bien en el año 2009 las redes sociales contaban con 830 millones de usuarios aproximadamente (ITU, 2010), un estudio reciente calcula que este número ascendió a 2.307 millones en 2016 (We Are Social, 2016). Twitter, una red que permite a sus miembros poner a disposición o enviar breves mensajes (de hasta 140 caracteres), se encuentra entre las redes sociales que lideran en número de usuarios a nivel global. En estos mensajes, conocidos como "Tweets" o trinos, los usuarios pueden dar a conocer al mundo lo que están haciendo o pensando, o bien comunicarlo a grupos de "listas" cerradas o a "twitterers" individuales (ITU, 2010).

Con lo anterior en mente, el uso oportuno de la información proveída por Twitter posibilitaría la identificación de usuarios, opiniones y tendencias respecto al tema de protección animal en Colombia y el mundo. Esto resulta primordial, ya que permite identificar a aquellos que están interesados en el tema y que constituyen población objetivo para responder a la encuesta digital que servirá como insumo para la elaboración del documento CONPES. Finalmente, el análisis de esta información permitiría desarrollar una estrategia para incentivar a esta población objetivo a que efectivamente responda la encuesta.

Metodología

Para llevar a cabo lo propuesto, se realizó una consulta a la Dirección de Justicia del Departamento Nacional de Planeación (DNP), en primera instancia. En dicha consulta, la dirección elaboró una lista con los usuarios de Twitter que potencialmente ejercen

---

[3] Estos insumos consisten en una serie de respuestas a preguntas sobre el tema. Dichas preguntas se encontrarán disponibles en una encuesta en formato digital.

mayor influencia en el tema de protección animal a nivel nacional. Luego, con la ayuda de GEPHI, un software que facilita el análisis de redes, se realizaron dos ejercicios con el fin de estudiar la cantidad de personas discutiendo sobre protección animal en Twitter. En ambos ejercicios se solicitó al programa que identificara los trinos que incluyeran alguna de las siguientes palabras clave: gatos, perros, adopción, protección y animales4. El programa permite recoger sólo un determinado número de trinos hacia el pasado, pero no impide recoger toda la información que se produce en el presente. De este modo, si el grupo de palabras clave ha aparecido en los trinos de manera muy frecuente, se tendrá mucha información sobre el tema en el pasado más reciente, pero poca en el pasado de más largo plazo. Por el contrario, si las palabras clave han sido mencionadas pocas veces en el pasado, se encontrarán más trinos de fechas más remotas.

La diferencia entre ambos consistió en que el primero restringía el número de usuarios a los listados por la Dirección de Justicia5, mientras en el segundo se tuvieron en cuenta todos los usuarios del mundo. El primer ejercicio se realizó el pasado lunes 20 de febrero entre las 9:00 am y las 3:00pm, mientras el segundo se realizó entre las 4:00 pm de ese mismo lunes y las 7:00 am del martes 21 de Febrero.

Dentro de las métricas utilizadas para valorar la centralidad de los usuarios en la red, se tienen:

- Distribución de grados tanto de entrada / salida:
  Se calcula el diámetro de los nodos con base en el número de conexiones entrantes que este tenga
- Modularidad:
  Aleatoriamente se selecciona un grupo de nodos para cuantificar la densidad de sus conexiones y contrastar dicha densidad con la de toda la red, si la densidad es mayor que la de toda la red se considera una comunidad
- Algoritmo de PageRank:
  Con base en el número de seguidores que tienen los usuarios que trinan, se asigna una calificación al usuario citado para determinar el diámetro de su nodo

Conclusiones preliminares

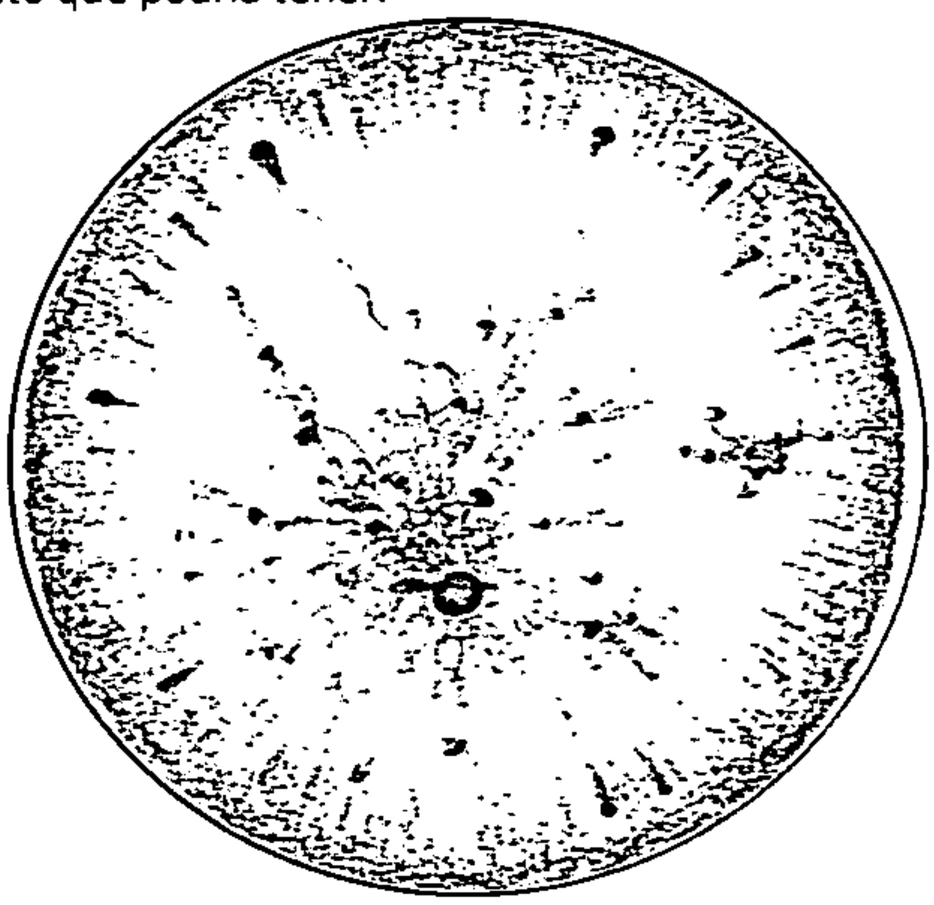
De ambos ejercicios se desprendieron lecciones diferentes. La más importante que deja el primero es la percepción de que el tema de protección animal es coyuntural en Colombia. Se trata de un tema sobre el cual personas pasivas reaccionan fuertemente frente a eventos específicos que generan polémica, resignación o rechazo. Esta primera impresión se deriva del comportamiento pasivo de los interesados en la temática en Twitter. Al analizar los *trinos* que alcanzó a recoger el programa en las 6 horas de duración del primer ejercicio, se encontró que estos databan de fechas muy remotas, indicando que no es un tema al cual se haga alusión frecuentemente. Finalmente, este

---

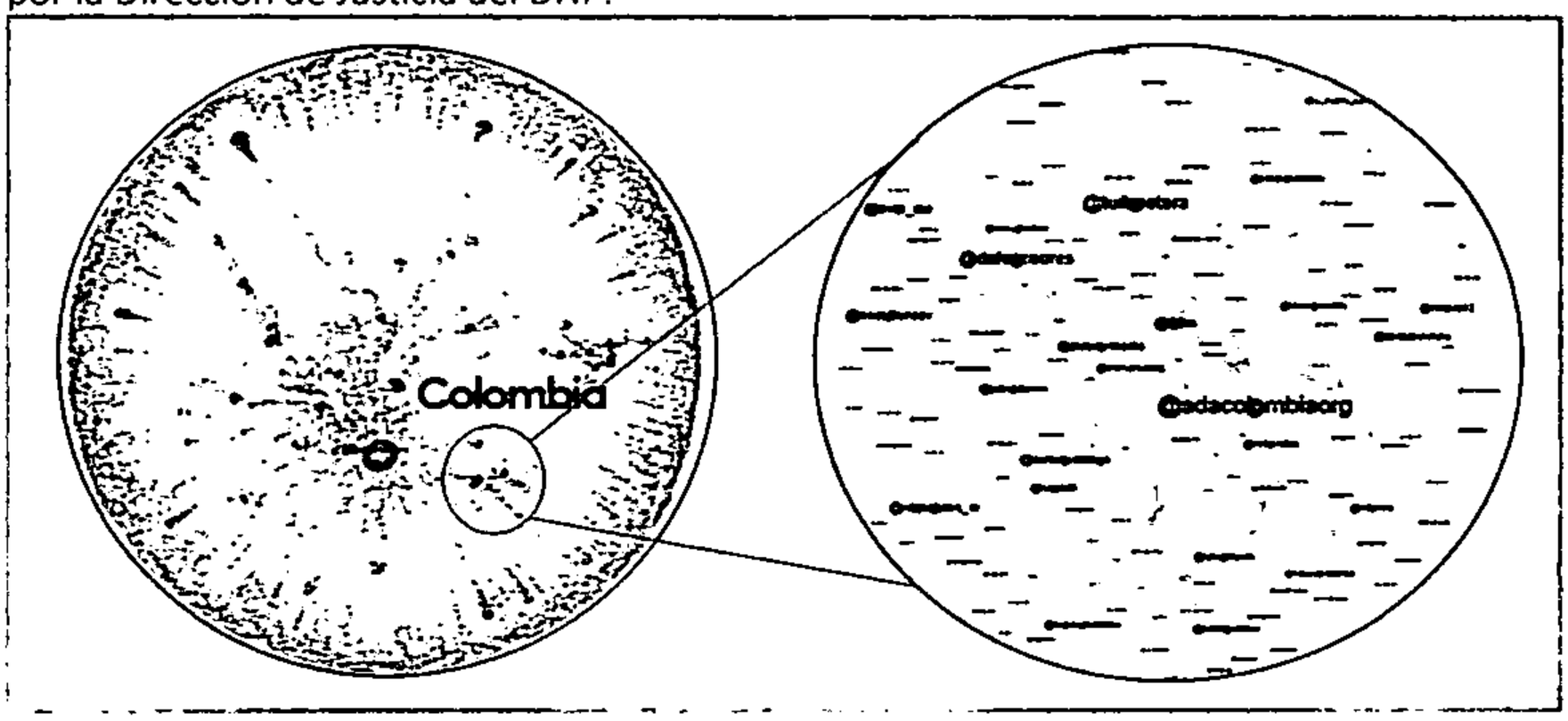
[4] Si bien es posible que alguna de estas palabras esté contenida en un mensaje que no esté relacionado con temas de adopción animal, el algoritmo que ejecuta el programa para encontrar los mensajes le permite relacionar las cinco palabras a la hora de hacerlo. Sin embargo, se realizó una limpieza de la base para evitar estos casos.

[5] Lo anterior no implica que el programa sólo identificara los *trinos* publicados por los usuarios de la lista, sino todos aquellos que tuviesen estrictamente alguna relación con las palabras y con dichos usuarios.

resultado sugiere que en Colombia el tema de protección animal no ha agotado todo el potencial movimiento que podría tener.

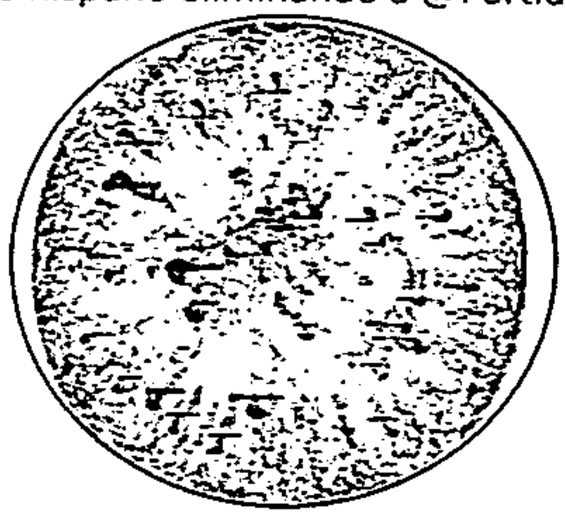


Ahora bien, poniendo el foco sobre los resultados del segundo ejercicio, es posible afirmar que: 1) la red de usuarios que ejercen mayor influencia en el tema de protección animal a nivel nacional se encuentran conectados a la red hispana líder en la temática, que; 2) hay muchos usuarios que, pese a interesarse por temas ambientalistas, no encuentran la protección animal como uno de sus temas de interés principal, y que; 3) los usuarios colombianos interesados en el tema se encuentran conectados a una red mundial de protección animal. De la información recogida por GEPHI en las 15 horas de duración de este ejercicio, fue posible identificar al Partido Animalista español –PACMA– como principal ente influyente en el tema de protección animal. Además, fue posible identificar un grupo de usuarios conectados con este partido y entre los cuales se encontraba gran parte de los usuarios de la lista proveída por la Dirección de Justicia del DNP.



Finalmente, considerando los resultados previamente expuestos se sugiere una estrategia para incentivar a los interesados en temas de protección animal a llenar la encuesta digital. Dicha estrategia consiste en promover la idea de inclusión social, permitiendo responder la encuesta a interesados en el tema de diferentes lugares del

mundo. Adicionalmente, se puede incentivar a estas personas a diligenciar la encuesta recordándoles que su respuesta los hace partícipes en la elaboración de un documento CONPES con alcance nacional.
Red de usuarios en el mundo hispano eliminando a @PartidoPACMA



Bibliografía:
Contagio Radio (2014) Protección animal es incluida en el PND 2014-2018 Recuperado de: http://www.contagioradio.com/proteccion-animal-es-incluida-en-el-pnd-2014-2018-articulo-7891/
ITU (2010) El incremento de las redes sociales, cambios en la web que conocemos. Recuperado de: http://itunews.itu.int/es/745-El-incremento-de-las-redes-sociales.note.aspx
ITU (2016) Measuring the information society report 2016. Recuperado de: https://www.itu.int/en/ITU-D/Statistics/Documents/publications/misr2016/MISR2016-w4.pdf
We Are Social (2016) Estudio sobre el estado de internet y las redes sociales en 2016. Febrero de 2016. Recuperado de: http://franbarquilla.com/estudio-sobre-el-estado-de-internet-y-las-redes-sociales-en-2016/

5. Análisis de Registros Únicos

Participantes: Edwin Torres, Juliana Ramírez y Juan Camilo Mejia. Bajo la coordinación de Mario Morales
El análisis realizado se llevó a cabo de acuerdo a las especificaciones dadas por el usuario, con el objetivo de extraer información relevante que evidenciara de forma clara y eficaz los niveles de duplicidad en la información capturada a través de los diferentes registros.
Usuario: Juliana Ramírez, Juan Camilo Mejía
Insumos: Encuesta explotación de datos, resultados encuesta explotación de datos
Producto: Matriz de campos Registros Únicos, Índices de redundancia de la información.
Supuestos: Información redundante, contenida en diferentes bases de datos (recursos redundantes), tiempos de diligenciamiento para el usuario.

Descripción

Se capturo información de los diferentes registros a través de búsquedas web y de la información suministrada a través de la encuesta de explotación de datos. Los siguientes datos resumen la información encontrada:

1. Entidades gubernamentales de orden nacional: 209
2. Entidades encuesta DNP: 152
3. Entidades que contestaron afirmativamente al registro único: 42
4. Entidades encontradas a través de búsqueda web: 21
5. Total entidades: 72
6. Total entidades con Registro Único (RU): 50
7. Total de RU con información de registro disponible en página web: 16
8. Total de RU incluidos en el análisis: 27 (*N*)
9. Total de campos únicos: 115 (*M*)

Características de los datos recopilados

- Confiabilidad baja, en ninguna captura se logró recopilar el total de campos del registro y en 4 solo se logró llegar a la información básica.
- La información de 32 URLs suministradas en la encuesta no corresponde a la ubicación del registro.
- Para el ingreso a un gran porcentaje del total de formularios encontrados, es necesario registrarse en el sistema correspondiente y validar a través de algún sistema de seguridad (ej. captcha) que el usuario es humano.
- No hay una diferenciación entre Sistemas de información (que se nutren de la información recopilada a través de los RU) y registros únicos.
- Hay una diferencia en el tamaño de la información capturada por los registros. Esta diferencia está en función del tipo de registro que se captura y el sector al que se encuentra dirigido: PN- Persona Natural, PJ-Persona Jurídica, Tramites, Documentos entre otros.

Análisis

Se construyó una matriz de *N* x *M*, en donde *N* corresponde al número de registros y *M* al número total de campos (Unión de los campos de todos los registros). A partir de la información contenida en la matriz se consideró que era necesario diseñar un grupo de indicadores que reflejaran de forma acertada y cuantificable los supuestos del proyecto.

Índices de redundancia

Se construyeron tres Índices de redundancia para determinar el nivel de repetición de la información contenida en la matriz: índice de redundancia del campo, índice de redundancia del registro y el índice de redundancia total. Estos índices toman valores entre [0, 1], en donde 0 corresponde a información no duplicada (Solo presente en una ocasión) y 1 corresponde a información duplicada en todos los escenarios en donde sea posible.

Como primer paso se diseñó una variable indicadora, z_{ij}, para determinar si el campo *j* está presente en el registro *i*. Se debe tener en cuenta que todo campo debe ser usado al menos en un registro.

$$z_{ij} = \begin{cases} 1, & \textit{si campo j esta presente en el registro i} \\ 0, & \textit{cualquier otro caso} \end{cases}$$

Índice de redundancia del campo (IRC)

Se define el Índice de redundancia del campo *j* sobre *N* registros como:

$$\sigma_j = \left(\frac{1}{N-1}\sum_i^N z_{ij}\right) - \frac{1}{N-1}$$

En donde $i = 0, 1 \ldots N - 1$
Este índice mide el nivel de repetición de un campo dentro del conjunto de registros.

Índice de redundancia del registro (IRR)
Se define el índice de redundancia del registro *i* como:

$$\gamma_i = \frac{1}{L_i}\sum_i^{L_i} \sigma_{ij}$$

En donde σ_{ij} corresponde al IRC de los campos contenidos en el registro *i*, y L_i corresponde al número total de campos contenidos en el registro *i*.
Este índice mide el nivel de repetición de un registro en función del IRC de los campos que contiene el registro.

Índice de redundancia total (IRT)
Se define el índice de redundancia total sobre un conjunto de *N* registros con *M* campos agregados como:

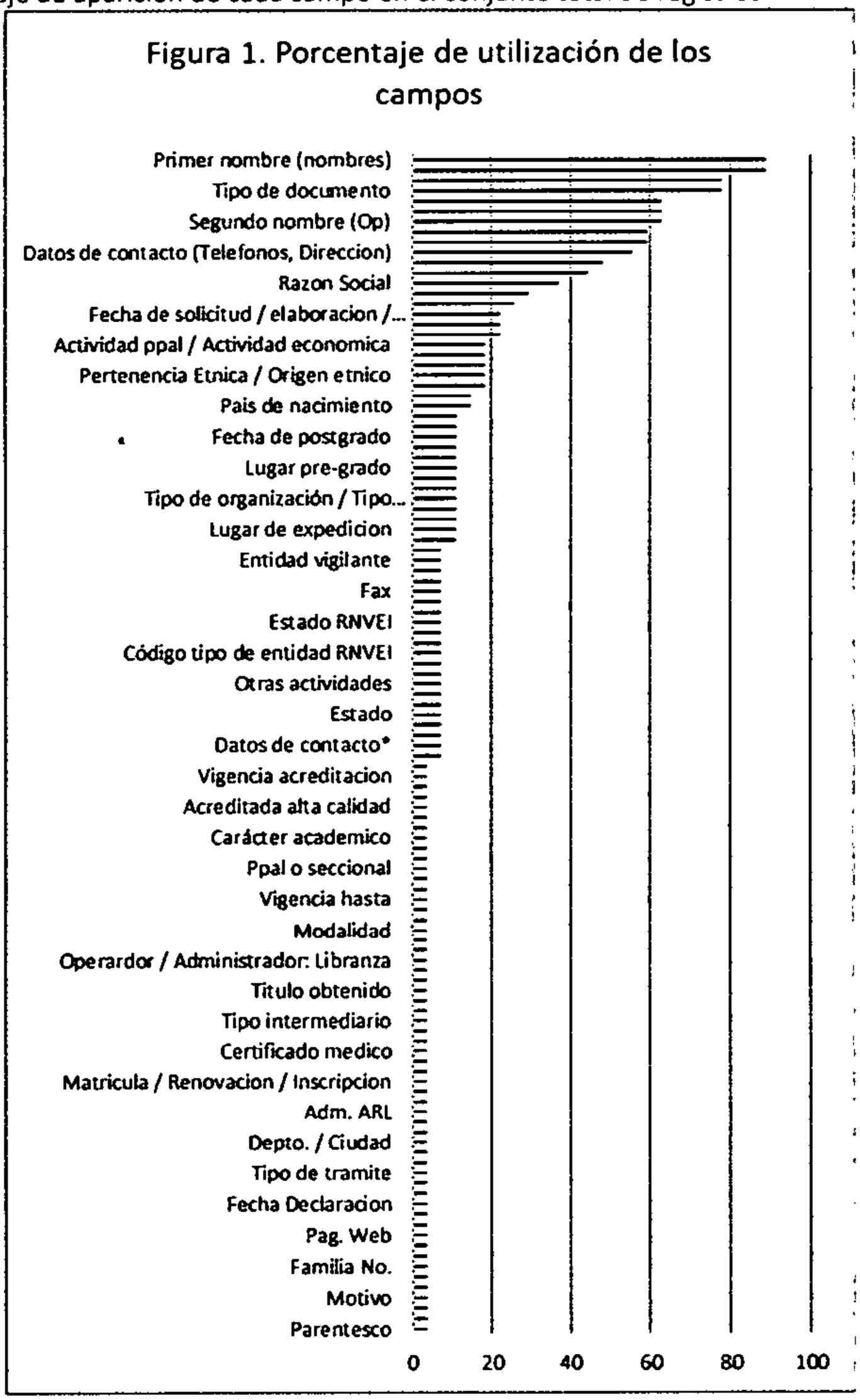
$$\varphi = \frac{1}{M}\sum_j^M \beta_j \sigma_j$$

En donde la variable β_j es utilizada para ponderar el campo *j* de acuerdo a un criterio específico. Por ejemplo, ponderar en función del espacio en memoria física que ocupa cada campo o ponderar en función de la relevancia del campo dentro del conjunto (es más relevante un campo con mayor IRC)

Resultados

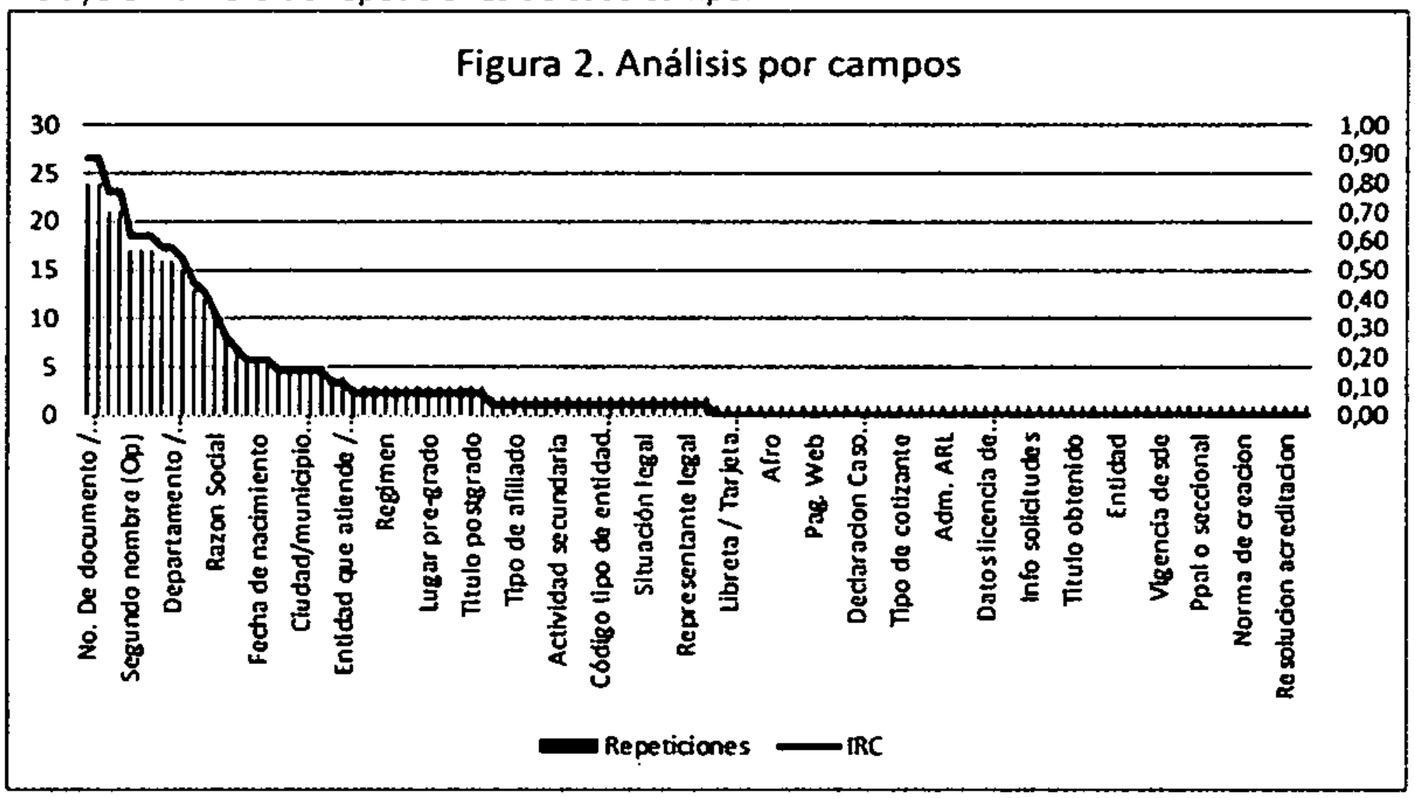
Haciendo un conteo por columnas se obtiene la Figura 1, en donde se muestra el porcentaje de aparición de cada campo en el conjunto total de registros.

Figura 1. Porcentaje de utilización de los campos



IRC - Índice redundancia del campo

Al realizar el cálculo del ICR, sobre los *M* campos, se obtienen los resultados de la figura 2., en donde los datos han sido organizados de acuerdo a su ICR. Adicionalmente, se incluye el número de repeticiones de cada campo.



IRR - Índice redundancia del registro

Al realizar el cálculo del IRR sobre los N registros se obtienen los resultados que se muestran en la Figura 2. Los registros están organizados de acuerdo al número de campos que contienen.

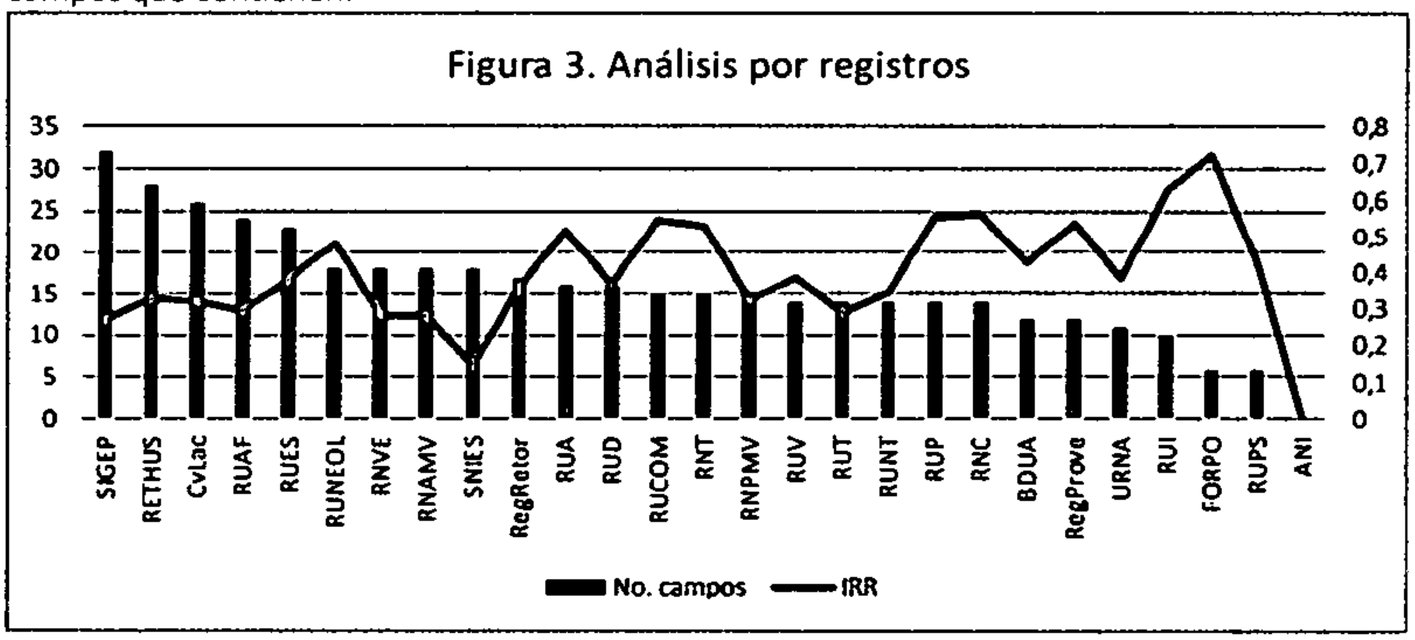


Tabla 2. IRR	
Registro	γ_i
FORPO	0,72
RUI	0,63

RNC	0,56
RUP	0,55
RUCOM	0,55
RegProve	0,54
RNT	0,53
RUA	0,52
RUNEOL	0,49
RUPS	0,43
BDUA	0,43
RUV	0,39
URNA	0,38
RUES	0,38
RUD	0,37
RegRetor	0,36
RUNT	0,34
RETHUS	0,33
RNPMV	0,33
CvLac	0,32
RUAF	0,29
RUT	0,29
RNVE	0,28
RNAMV	0,28
SIGEP	0,27
SNIES	0,15
ANI	0

IRT - Índice redundancia total

Al realizar el cálculo del IRT se obtiene,

$\boldsymbol{\varphi = 0{,}104}$

Se utilizó un $\beta_j = {}^{1}/_{115}$, que corresponde a una ponderación uniforme de los campos.

Conclusiones

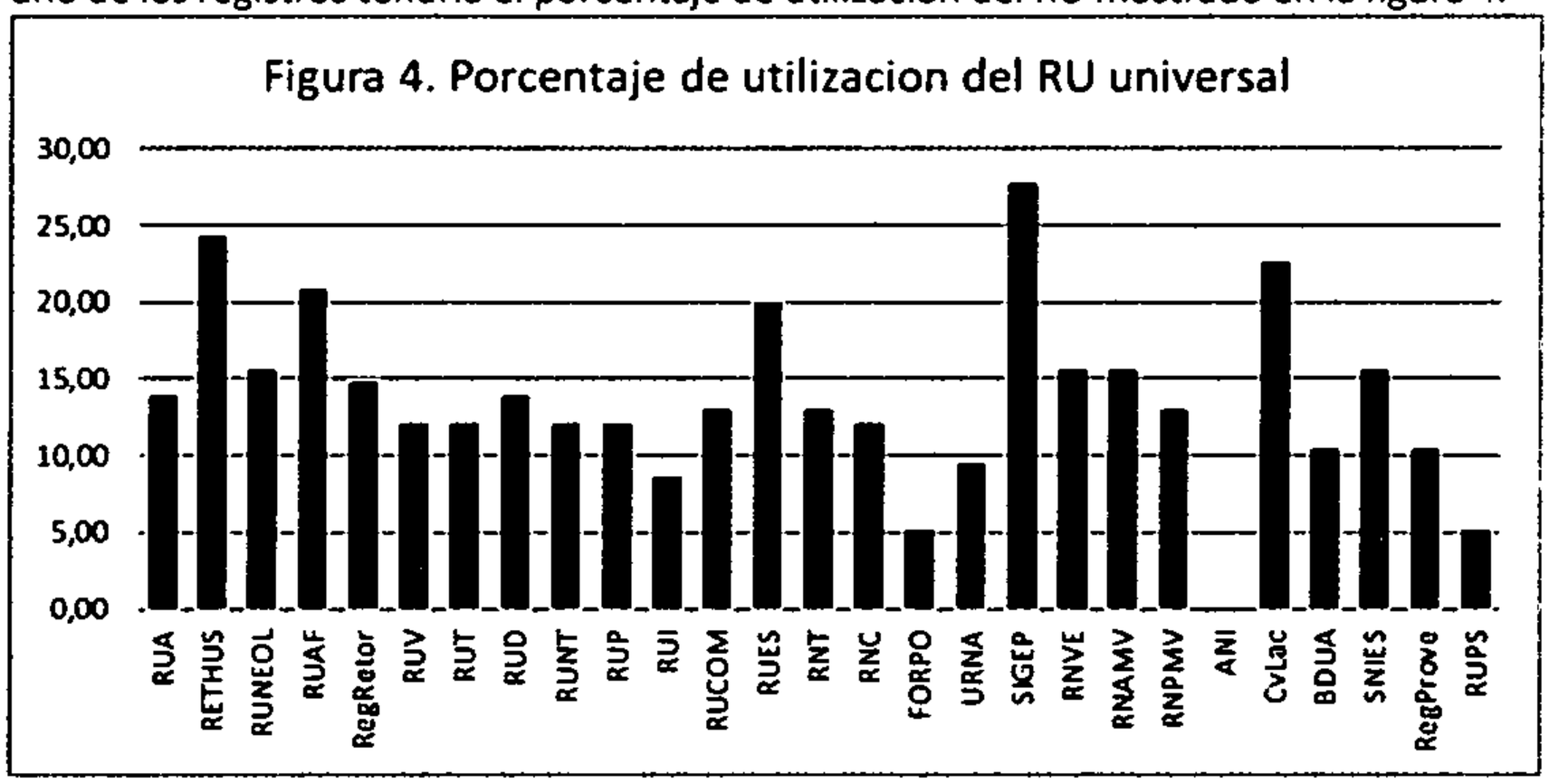
Un grupo de 10 campos tienen un IRC superior a 0,5. Lo que evidencia la existencia de una información que es altamente relevante en las instituciones para "indexar" los registros capturados.

Sin embargo, también existe un numero de campos (105) con un IRC inferior a 0.5, lo que muestra el nivel de especificidad de gran parte de la información recopilada en los registros.

A nivel de registros se observa que todos los registros poseen información redundante en por lo menos un registro del conjunto. En especial 8 registros tienen IRR superior a 0,5 lo que indica que el registro tiene redundancia de la información con más de la mitad de los registros del conjunto.

Sin embargo, se encuentra que el IRT es bajo, lo que sugiere que una gran cantidad de los registro poseen campos únicos.

Asumiendo un Registro Universal - RU, aquel que contiene un total de *M* campos, cada uno de los registros tendría el porcentaje de utilización del RU mostrado en la figura 4.



6. Proyecto Big Data para el monitoreo de precios agropecuarios

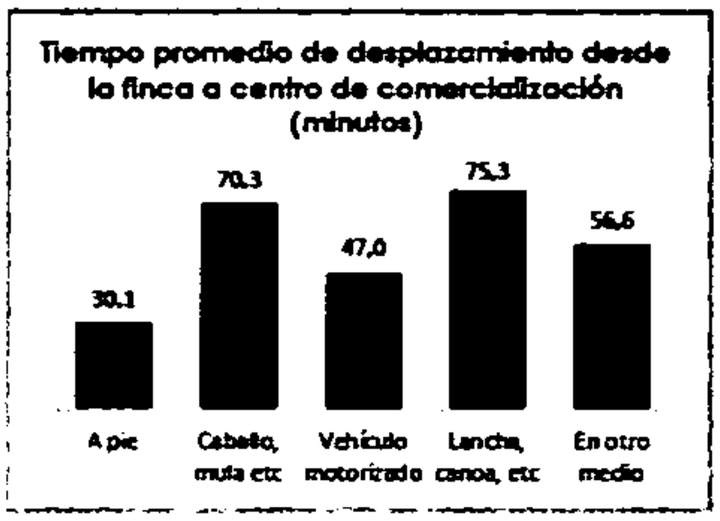
Participantes: Daniel Rojas. Bajo la coordinación de Mario Morales.

Justificación

De la mano de la Dirección de Desarrollo Rural Sostenible (DDRS) del DNP, se identificó que Los productores agropecuarios se enfrentan a barreras relacionadas con el desconocimiento del precio de compra del consumidor final y altos tiempos de desplazamiento a centros de comercialización, lo que genera:

- Bajo poder de negociación de los productores con los intermediarios.
- Alta vulnerabilidad a choques en dinámica del mercado.

En donde el 76,52% de las unidades de producción con cultivos comercian parte de su producción según cálculos realizados con información obtenida por el Consejo Nacional Agropecuario (2014)



La DDRS-DNP y el MADR carecen de información histórica (diaria) de fácil acceso y consumo sobre el comportamiento de los precios de los productos agropecuarios. Razón por la cual, consultar diariamente lo publicado por el portal SIPSA del DANE y generar una base de datos histórica de la información brindaría un insumo adicional para las entidades mencionadas.

Objetivos

El objetivo general del proyecto es entonces, desarrollar una herramienta que permita monitorear el comportamiento de los precios de los productos agropecuarios comercializados en las centrales mayoristas del país, según información del DANE.

Para llevar a cabo este objetivo se establecieron 3 objetivos específicos:

- Construir una base de datos histórica, de fácil acceso y consumo que recopile la información publicada diariamente por el DANE.
- Predecir el comportamiento de los precios de los productos agrícolas comercializados en las centrales mayoristas, a partir del mejor modelo de ajuste.
- Visualizar patrones de relación entre las series precios mediante mapas de calor.

Análisis exploratorio de los datos

Características de los datos

La Fuente de información para el desarrollo de este Proyecto es el portal SIPSA del DANE, en donde se hace pública la información de los precios agropecuarios en las principales centrales mayoristas del país, así como la variación porcentual respecto al día anterior.

Las principales características de la información de este portal son:

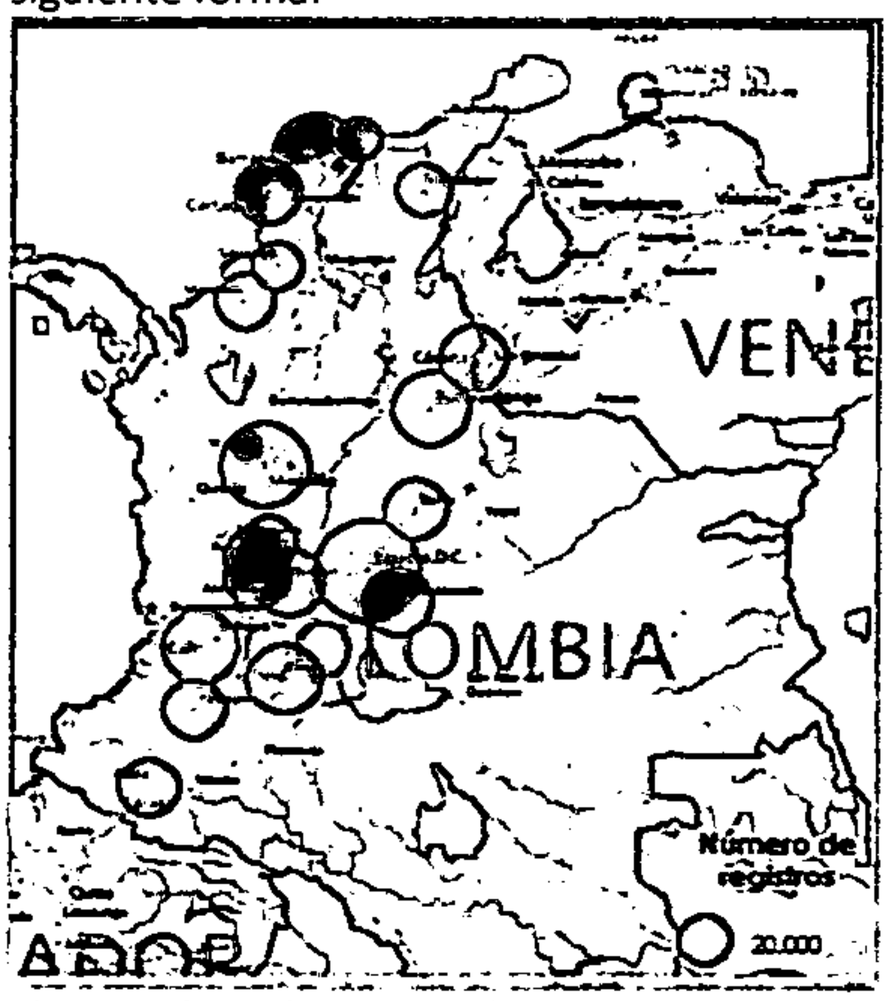
1. El DANE publica diariamente un archivo Excel con información del precio por kilogramo (y su variación) de hortalizas y verduras, frutas frescas, y tubérculos y plátanos.

2. Los precios solo se publican para días hábiles.

3. Existen errores de digitación o cálculo en la información recopilada.

4. Se presenta asimetría en la información reportada por productos y centrales mayoristas.

En promedio la base de datos crece diariamente 417 registros, su ubicación geográfica está distribuida de la siguiente forma:



La cantidad de datos diarios al 20 de septiembre son 1.126 el top 10 de centrales con mayor número de registros son:

Ciudad y central	Número de registros
1. Bogotá, Corabastos	40.106
2. Medellín, CMA	40.052
3. Pereira, Mercasa	30.724
4. Cúcuta, Cenabastos	25.134
5. Bucaramanga, Centroabastos	25.042

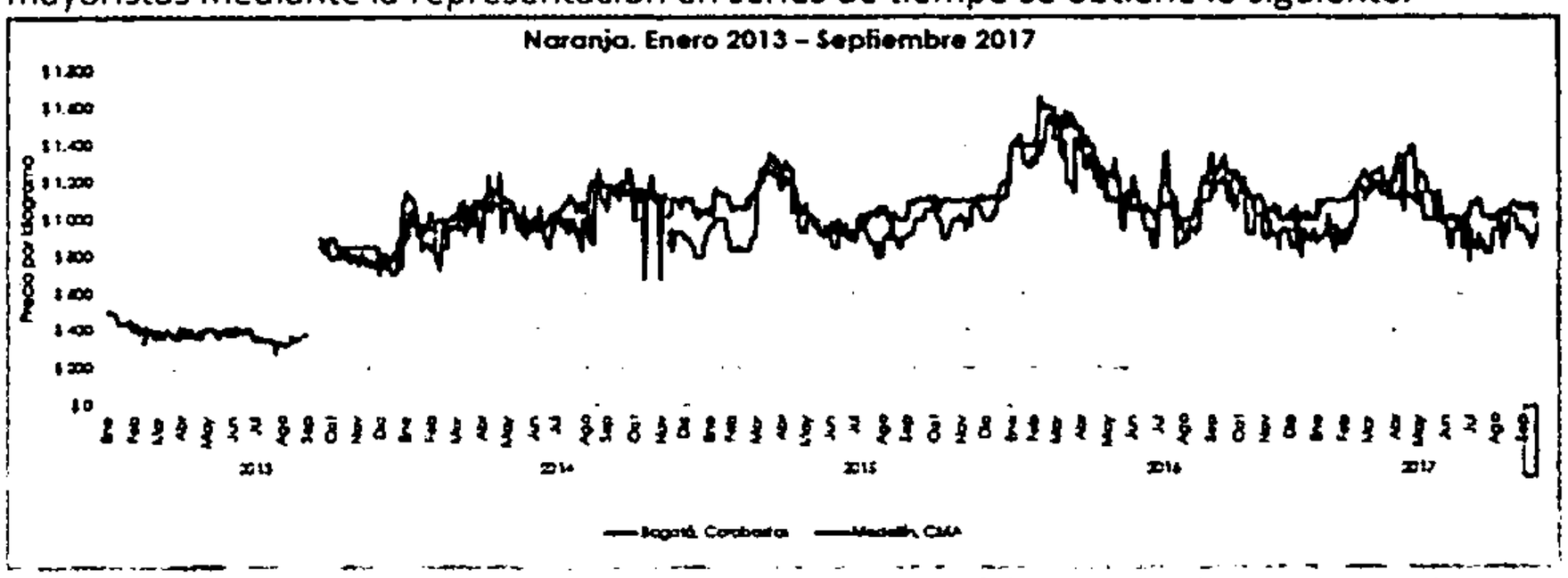
6. Armenia, Mercar	23.508
7. Barranquilla, Granabastos	23.206
8. Cartagena, Bazurto	22.901
9. Tunja	21.272
10. Neiva, Surabastos	21.159

La información capturada desde el 2 de enero de 2013 hasta el 11 de septiembre de 2017 para los días hábiles solamente cuenta con 22 centrales mayoristas y 36 productos agrícolas:

Ciudad y central	Porcentaje de productos
Medellín, CMA	97%
Armenia, Mercar	94%
Bogotá, Corabastos	94%
Bucaramanga, Centroabastos	94%
Tunja	94%
Ibagué, Plaza La 21	89%
Pereira, Mercasa	86%
Cúcuta, Cenabastos	83%
Manizales	83%
Popayán	83%
Cali, Cavasa	81%

Ciudad y central	Porcentaje de productos
Pereira, La 41	81%
Villavicencio, CAV	81%
Barranquilla, Granabastos	78%
Cali, Santa Helena	78%
Montería	78%
Cartagena, Bazurto	72%
Valledupar, Mercabastos	72%
Neiva, Surabastos	69%
Pasto, El Potrerillo	69%
Sincelejo	58%
Santa Marta	42%

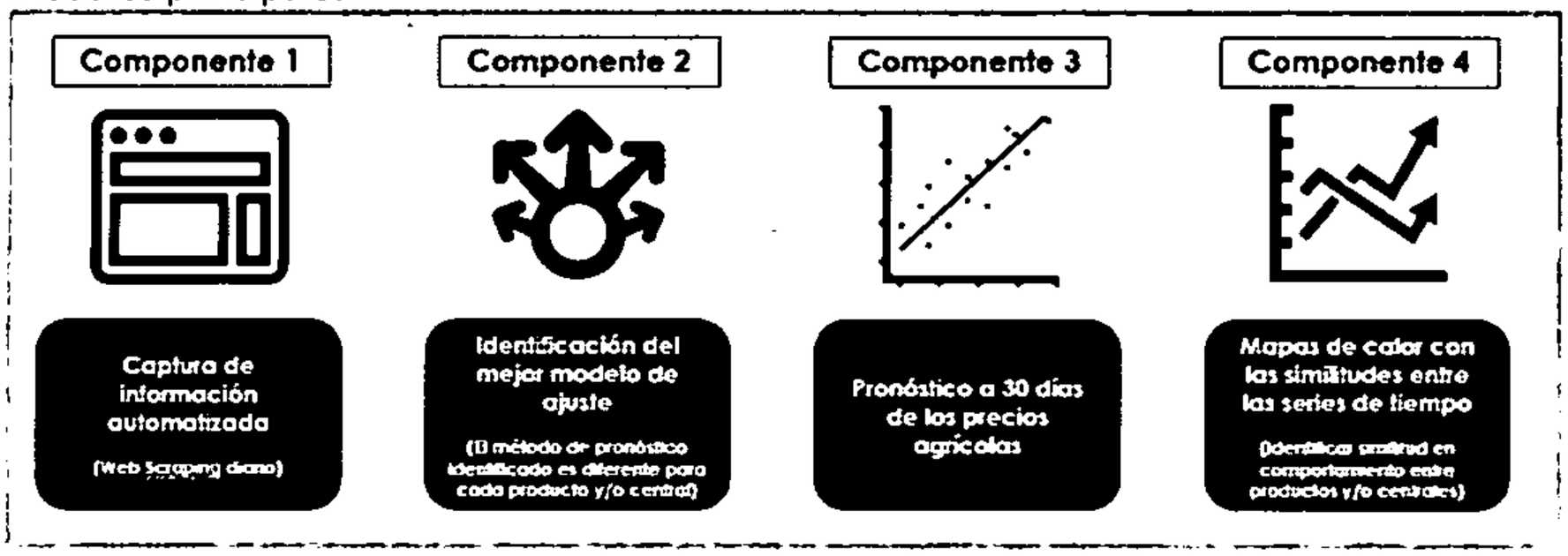
En cambio, analizando el porcentaje de centrales por producto se tiene:
Cuando se selecciona cualquier producto y se contrastan los datos entre centrales mayoristas mediante la representación en series de tiempo se obtiene lo siguiente:



Se pueden evidenciar productos que tienen datos faltantes, atípicos y cambios estructurales en su comportamiento.

Componentes desarrollados en el proyecto

Dada la naturaleza de la información que se está tratando y el fin general que es crear una herramienta para monitorear dichos precios, se determinó realizar el proyecto en 4 módulos principales:



Componente 1

Para este módulo se desarrolló un algoritmo que estuviese en la capacidad de determinar cuáles fechas son faltantes en la base de datos histórica y fuese directamente a la página del DANE y bajase los archivos correspondientes a ese vector de fechas faltantes.

Una vez se encuentran descargados los archivos, el algoritmo genera un ciclo donde va buscando por cada archivo el nombre de la central mayorista, el producto y su correspondiente precio en kg para luego adicionar a la base de datos principal.

Como resultado final se obtiene una base de datos long-term que guarda la información de los precios diarios para 740 posibles combinaciones de productos y centrales.

Componente 2

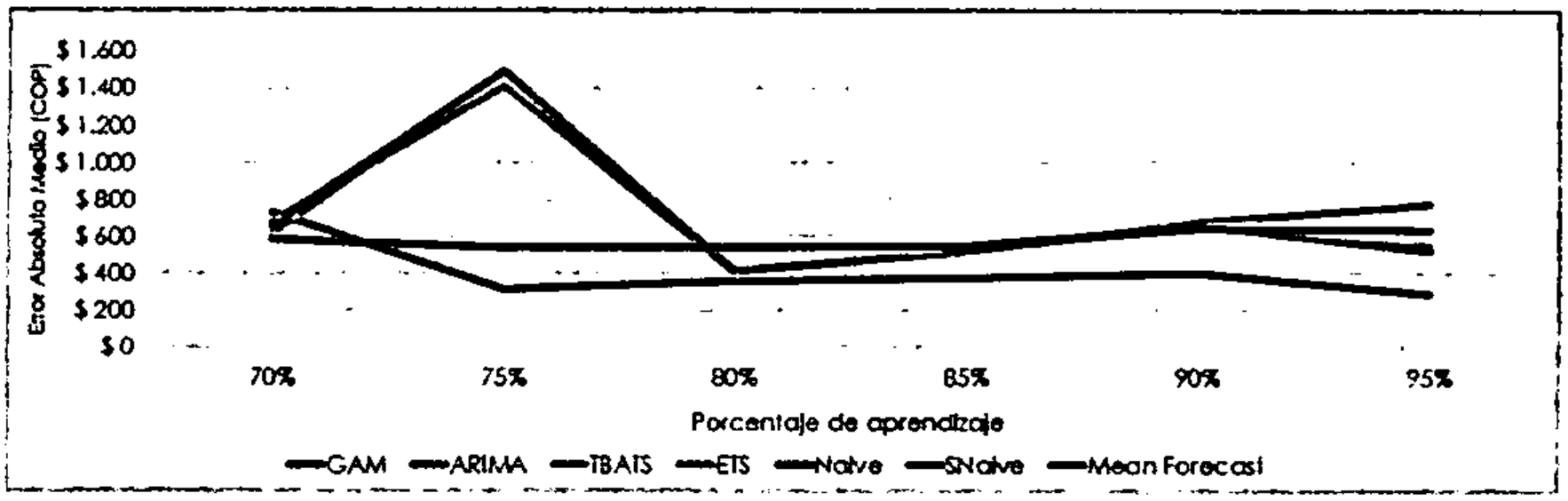
Para este módulo se desarrolló un algoritmo que evaluara 6 diferentes tipos de regresiones para series de tiempo que según el paper *Forecasting at Scale cuyos autores son: Sean J. Taylor – Benjamin Letham los cuales fueron:*

- TBATS model
- ARIMA
- Error, trend, seasonality
- Naive
- Generalized Additive Model
- Mean Forecast

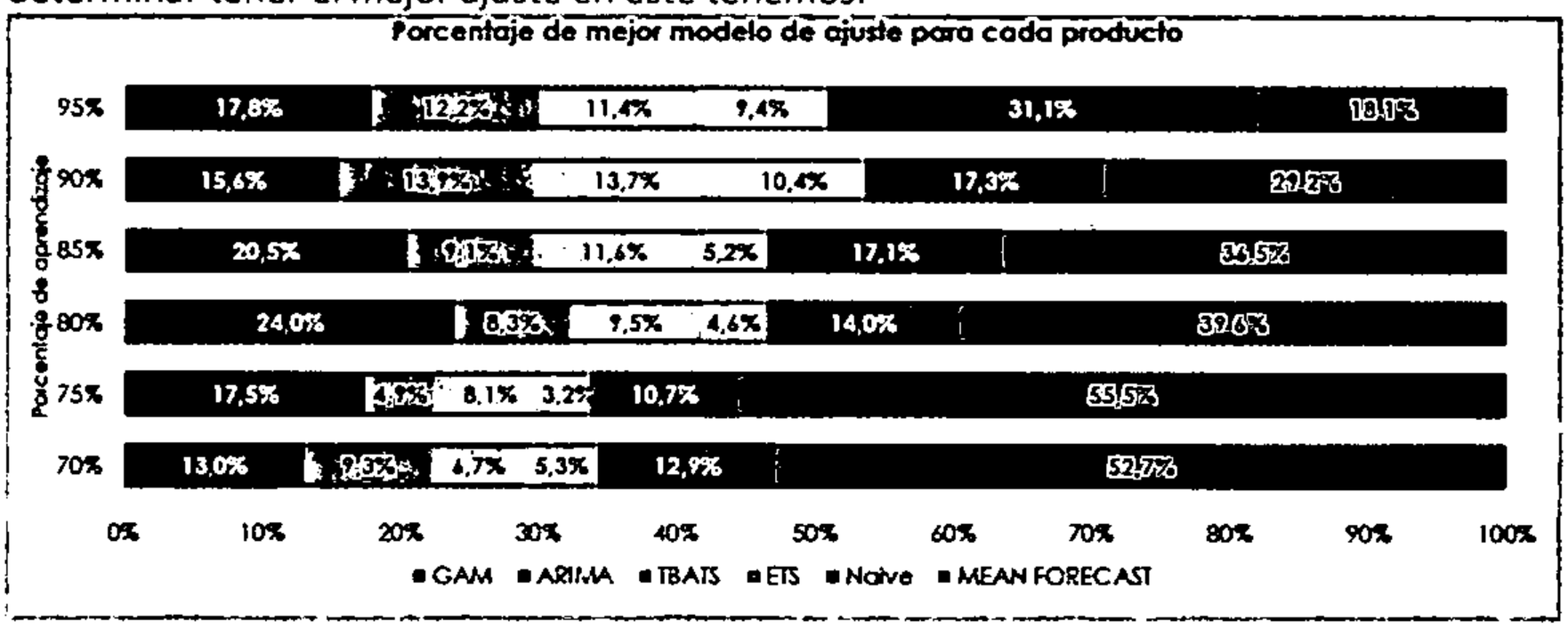
El motivo de evaluación con diferentes modelos es básicamente determinar para cada serie de tiempo disponible (producto X central) cual es el mejor ajuste basado en la comparación de los errores absolutos medios de cada modelo y variando el índice de aprendizaje del módulo de 5% en 5%.

El índice de aprendizaje es un porcentaje de sensibilidad que permite al modelo ver únicamente una parte de toda la serie en cuestión, de esta forma se obtendrá un set de datos complementarios para realizar el pronóstico y contrastarlo con los datos observados y de esa manera ver que tan bien son las predicciones realizadas.

El resultado para un producto aleatorio al graficar la diferencia de errores ante diferentes porcentajes de sensibilidad es:



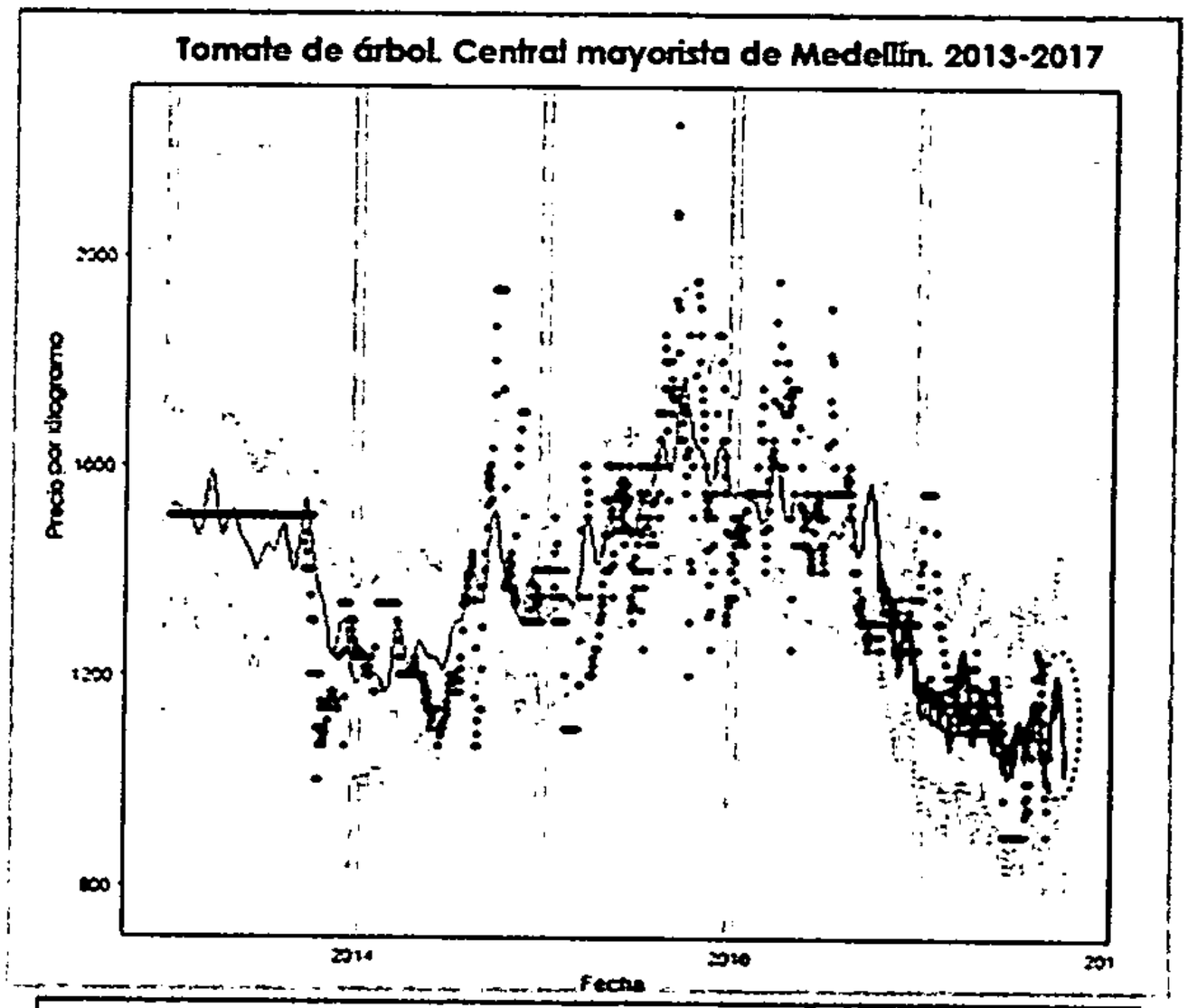
Luego, si este procedimiento se realiza para cada posible combinación de productos y se determina la distribución de cada modelo según el número de productos que determinar tener el mejor ajuste en este tenemos:

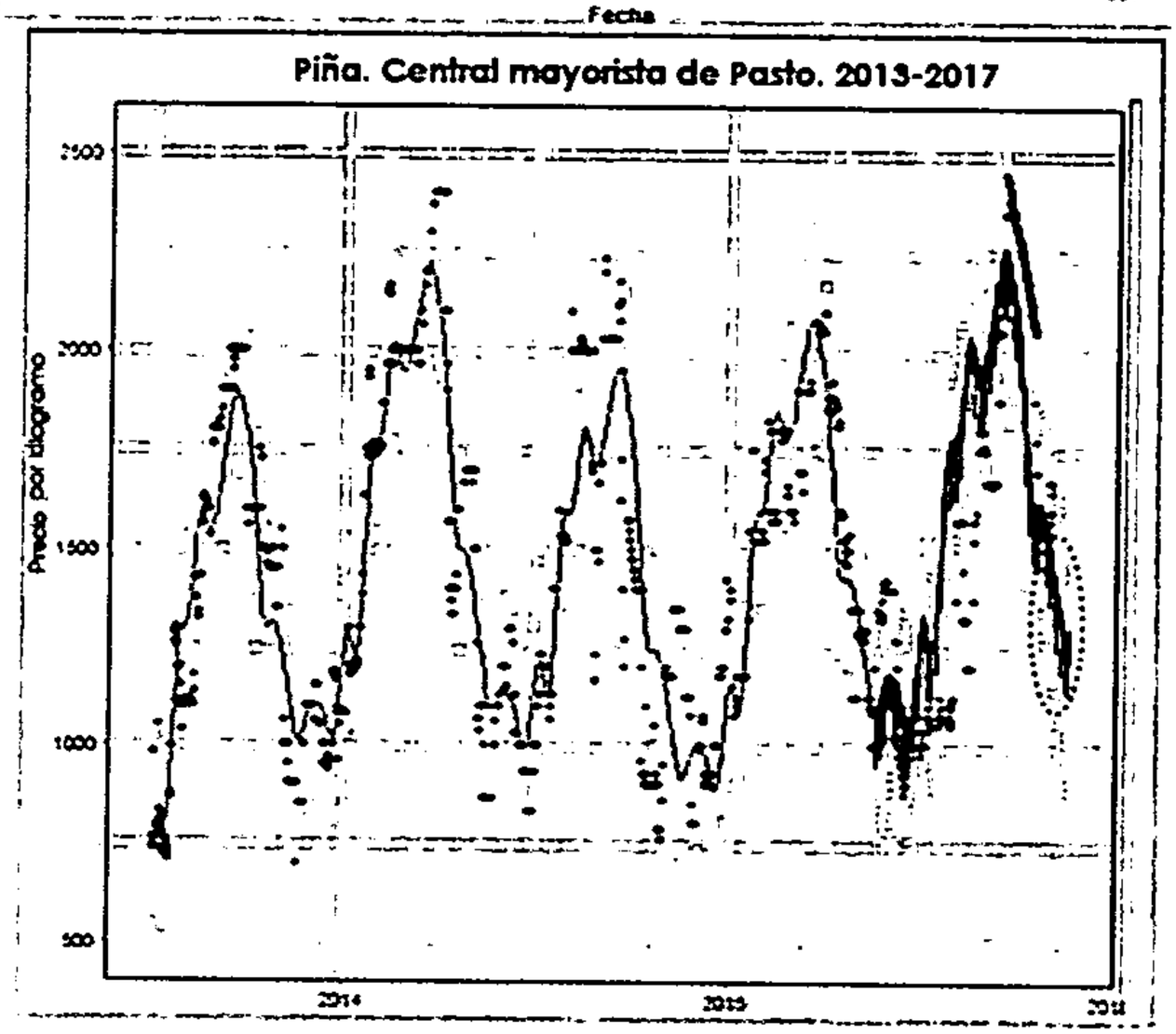


Finalmente, el algoritmo está en la capacidad de generar una tabla de "Algoritmo Ganador" para cada producto y que posteriormente sea utilizado por el componente de predicción (componente 3)

Componente 3

Luego de poder determinar cuál es el mejor modelo de ajuste se procede a realizar la predicción para los próximos 30 días contando desde la última fecha que se tiene registro en la base de datos. Para ello el algoritmo verifica que modelo aplicar y luego de generar las estimaciones necesarias procede a exportar una imagen formato png con el resultado de la predicción delineado con un tono azul, mientras que la parte de aprendizaje o la parte visible por el modelo con el cual se utilizó para realizar dicha predicción tiene una tonalidad roja. A continuación, algunos ejemplos:


Tomate de árbol. Central mayorista de Medellín. 2013-2017
Precio por kilogramo
2000
1600
1200
800
2014
2016
Fecha


Piña. Central mayorista de Pasto. 2013-2017
Precio por kilogramo
2500
2000
1500
1000
500
2014
2015

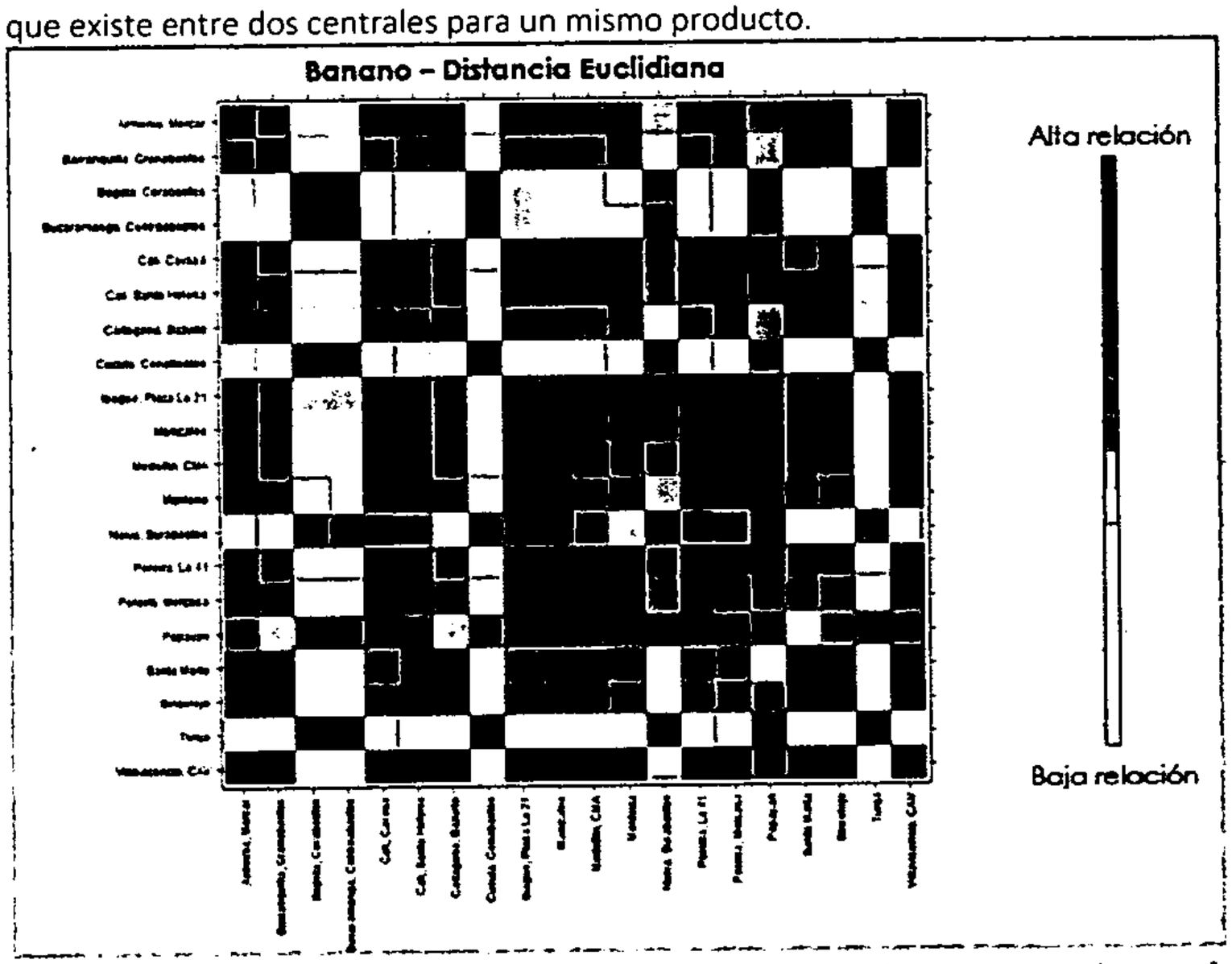
Componente 4

Para demostrar que la información capturada en el componente 1 puede ser de gran utilidad aparte de predecir el comportamiento de los precios en el mercado, se adicionó este componente buscando comparar el comportamiento de un producto en diferentes centrales mayoristas, logrando identificar un comportamiento geográfico crucial para la determinación de los precios en las zonas representadas por las centrales mayoristas.

Para este análisis se utilizaron 4 pruebas vectoriales distintas:

- Euclidian Distance
- Manhattan Distance
- Minkowski Distance
- Infinitive Norm Distance

Los resultados son imágenes png con mapas de calor determinando el nivel de cercanía que existe entre dos centrales para un mismo producto.



Los principales resultados de este módulo fueron encaminados a que visualmente fuese mucho más sencillo consumir las distancias calculadas, por tal razón el algoritmo está diseñado para correrse una vez y de volverse a ejecutar se esperaría encontrar cambios si han ocurrido choques externos que hayan alterado el comportamiento de los precios hacia el pasado.

Acceso al código de desarrollado

La herramienta se desarrolló en el lenguaje de programación R versión 3.3.3 y el código fuente de todo el proyecto se encuentra en el siguiente link:

- https://github.com/rojasdaniel/Monitoreo-de-precios-AGRO/tree/codigo_final

También se encuentra corriendo en el servidor VanaliticaDev del DNP el código para capturar información diaria desde el 1 de septiembre de 2017 actualizándose a las 7:00 p.m. y enviando la información a una carpeta compartida en OneDrive con todos los involucrados en el proyecto.

7. Modelo para la clasificación de proyectos de regalías

Participantes: Edwin Torres y José Zea. Bajo la coordinación de Mario Morales

Problema

Diseñar un sistema que permita evaluar, en la etapa de formulación, el éxito o el fracaso de un proyecto de inversión presentado al Sistema General de Regalías. Este modelo debe realizar un análisis de la información de texto que se incluye en los estudios técnicos que describen a cada proyecto.

Insumos

Documentos formulación proyectos de regalías (pdfs), MGA, SUIFP

Objetivo

Apoyar en el acopio y depuración de la información para construir un modelo que permita predecir si un proyecto de regalías presentará una alerta:

1. Lectura de base de datos de regalías del servidor *vbidev* de la base de datos MAPAI_Fuente1.

2. Lectura de las siguientes variables de texto:

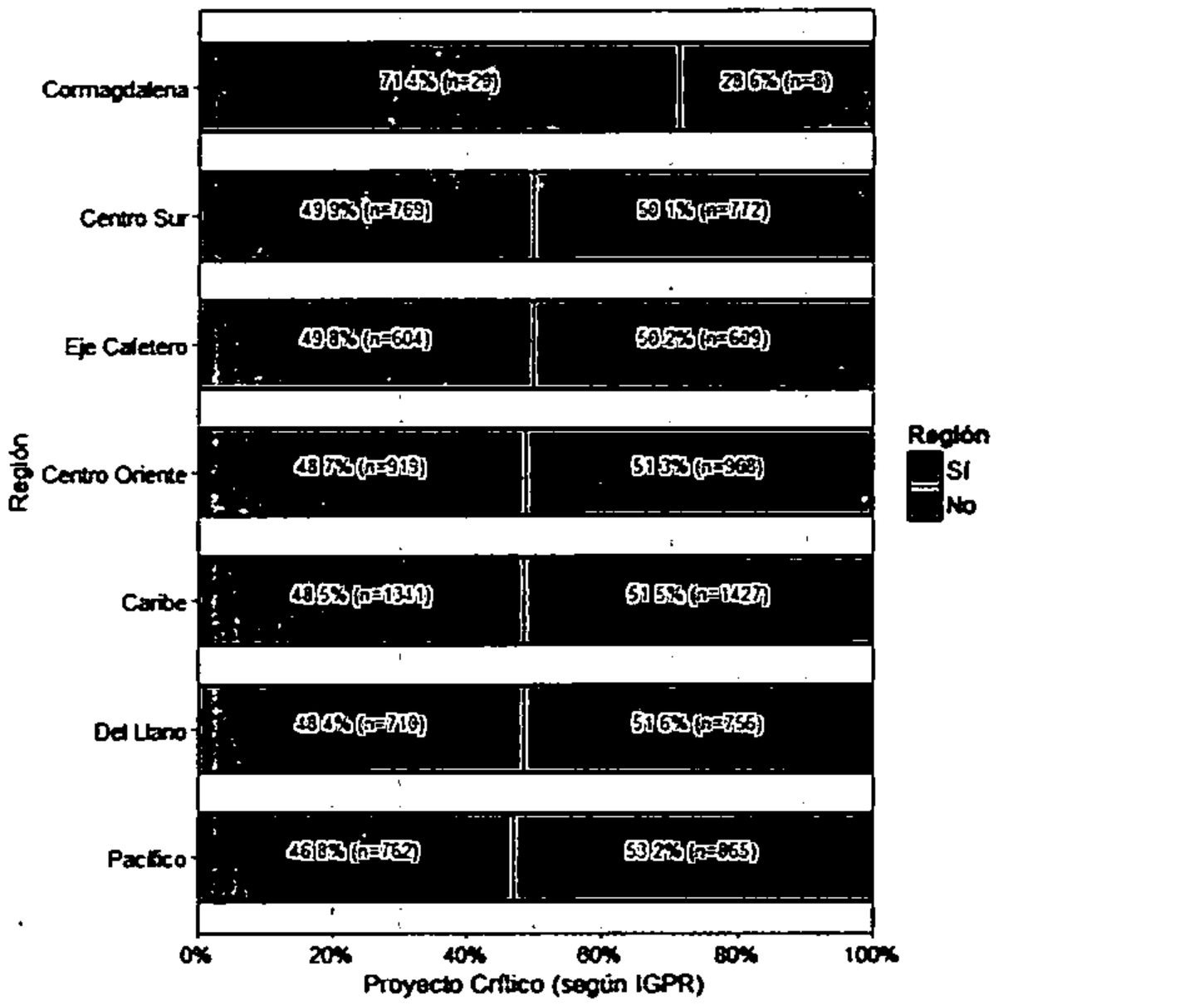
    - Identificador del texto: BPIN,
    - Nombre del proyecto
    - Problema central
    - Descripción
    - Causas directas e indirectas
    - Efectos directos e indirectos
    - Objetivo general
    - Objetivos específicos.
    - Actividades
    - Productos

    a. Limpieza básica del texto: eliminación de símbolos de puntuaciones, números, espacios
    b. Aplicación de correctores ortográficos
    c. Lematización con *Treetagger*
    d. Remover nombres geográficos del texto.
    e. Corregir las palabras mal lematizadas por *Treetagger*

3. Lectura y procesamiento de índice de Gestión de Regalías

4. Análisis exploratorio de datos.

    Se analizó el índice de regalías por



Descripción

No hay un mecanismo que permita identificar anticipadamente (en la etapa de formulación) si un proyecto entrará en estado crítico o no durante su etapa de ejecución.

El siguiente documento describe el proceso realizado para el diseño y construcción de la solución al problema formulado. El proyecto se desarrolló en tres fases: la primera en la cual se contextualiza el proceso que genera la información a analizar, en concreto a la metodología empleada en la formulación, aprobación y seguimiento que adelanta el Sistema General de Regalías (SGR) a los proyectos que solicitan su financiación. En la segunda fase se realiza un análisis exploratorio de los datos y un preprocesamiento de la información de texto recopilada y su transformación en un tipo de dato apto para ser suministrado al modelo. En la fase tres se implementan algoritmos de clasificación que permitan determinar la probabilidad de ocurrencia de una alerta.

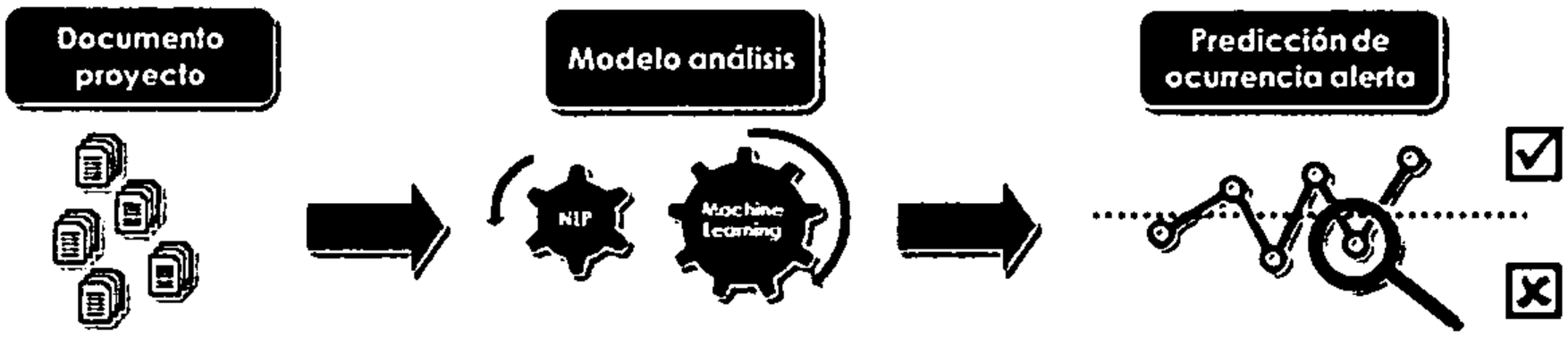


Figura 5. Flujo para la predicción de alertas.

Datos

La consolidación de la base de datos utilizada se realizó mediante la captura de la información de los proyectos almacenada en los sistemas de información usados para

tal fin. Estos son la MGA, SUIFP y Gesproy. Se cuenta con un total de 10688 proyectos desde el año 2012.

A partir de este conjunto se hace una revisión de las variables disponibles, tomando en cuenta que estas deben estar presentes para todos los proyectos. De esta revisión se define que se trabajará con las variables contenidas en los campos de la las fichas EBI (Estadísticas básicas de inversión), fichas en las cuales se encuentra la información básica de los proyectos.

La ficha EBI se encuentra estructurada por los siguientes módulos:

1. Módulo De identificación del problema o necesidad
2. Módulo de preparación de la alternativa de solución
3. Módulo de evaluación de la alternativa de solución
4. Módulo de programación
5. Módulo de decisión

La extracción de la información de texto relevante se hace del módulo 1, en el que se encuentran los siguientes campos:

- Problema central
- Descripción de la situación existente
- Magnitud actual
- Causas que generan el problema (Dir e ind)
- Efectos generados por el problema (Dir e ind)
- Objetivo general
- Objetivos específicos
- Alternativa de solución

Cada campo de estos corresponde a un texto de longitud variable. Debido a la forma en la que la información se ingresó a las bases de datos fue necesario pasar todos los textos por un corrector ortográfico que incluía terminología propia de las diferentes regiones del país. Posteriormente se procesó el texto a través de un lematizador con el fin de reducir el vocabulario sin afectar el significado de las palabras dentro del contexto de cada campo. El resultado de este procesamiento se resume en lo siguiente:

- 10.529 Proyectos
- 1'591.595 palabras (sin stop words y lematizado)
- 68.539 palabras únicas
- Longitud promedio 151 palabras por documento
- Promedio 91 palabras únicas por documento

Para las etiquetas se trabajó con el índice de gestión de regalías IGRP, que es un indicador para medir y valorar la gestión de las entidades ejecutoras que visibiliza la gestión de los proyectos. Esta herramienta hace un análisis contemplando dos dimensiones:

- Administrativa
    - Transparencia y medidas del SMSC
- Desempeño
    - Eficiencia y eficacia

En resumen, el IGRP es una sumatoria de los factores ponderados:

IGPR = f (Transparencia, SMSCE, Eficiencia, Eficacia | θ)
Se definió un umbral ρ con el objetivo de segmentar el conjunto de datos en dos categorías.

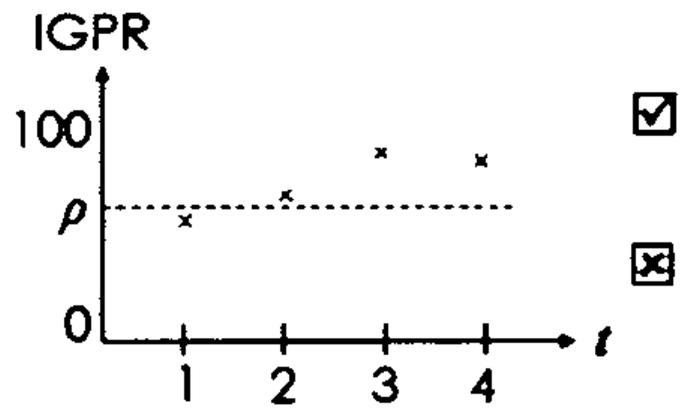


Figura 6. Umbral sobre el IGPR para la generación de etiquetas.

$$y(IGPR) = \begin{cases} 1, & IGPR \leq \rho\ ,\ \forall\ t \\ 0, & IGPR > \rho\ ,\ \forall\ t \end{cases}$$

Transformación del texto

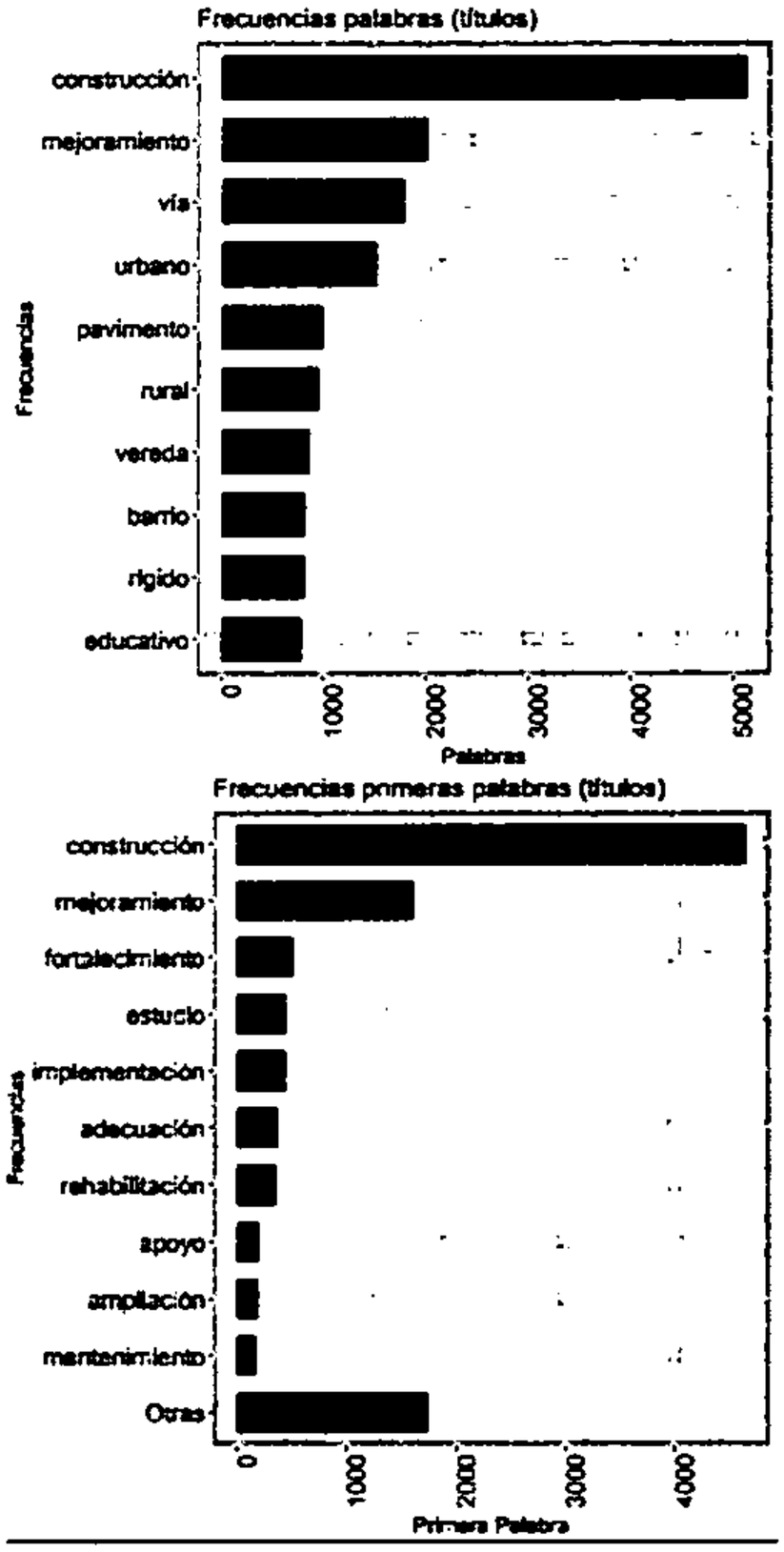
Se inicia con un pre-procesamiento estándar del texto para reducir los posibles errores que se hayan generado en la etapa de ingreso y captura de la información por parte de los proponentes de los proyectos. Este preprocesamiento consiste en los siguientes pasos:

- Segmentación por palaras para todo el texto
- Eliminar información no textual (caracteres especiales, puntuación, símbolos, palabras con longitud menor a dos caracteres)
- Lematizar todo el texto
- Corrección ortográfica.

Se continuo con una etapa de procesamiento específico, es decir, en función del campo a analizar (SGR), fuertemente influenciado por suposiciones que debemos hacer para segmentar eficientemente el análisis.

1. Títulos: Hipótesis - Primera palabra es el indicador más importante para determinar en que consiste el proyecto.
2. Cadena de valor
    a. Objetivos (General, específico): Primeras palabras
    b. Producto
    c. Actividades

Los resultados de esta etapa se pueden apreciar con mayor claridad en las siguientes graficas:


Frecuencias palabras (títulos)
construcción
mejoramiento
vía
urbano
pavimento
rural
vereda
barrio
rígido
educativo
Frecuencias
0
1000
2000
3000
4000
5000
Palabras
Frecuencias primeras palabras (títulos)
construcción
mejoramiento
fortalecimiento
estudio
implementación
adecuación
rehabilitación
apoyo
ampliación
mantenimiento
Otras
Frecuencias
0
1000
2000
3000
4000
Primera Palabra

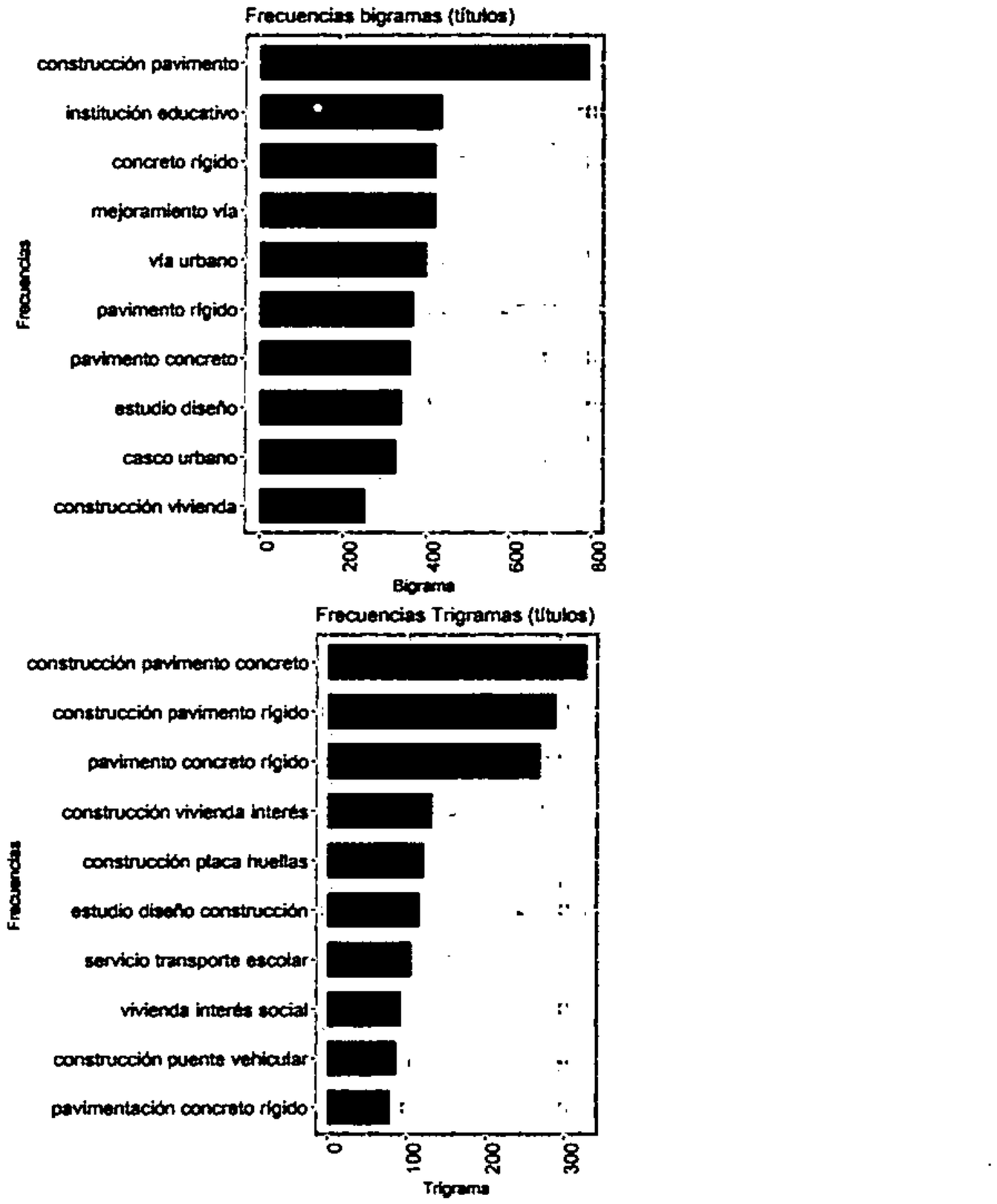


Figura 7. Resultados del análisis de frecuencias de palabras y n-gramas

Vectorización

Para convertir el texto en información cuantificable que pueda ser ingresada al modelo clasificador, fue necesario realizar un proceso de vectorización sobre el texto completo de cada proyecto. De esta forma cada proyecto es representado por un vector M-dimensional.

Se utilizo la vectorización a través de redes neuronales implementada a través del algoritmo Doc2Vec. El resultado de esta vectorización se representa en la siguiente figura a través de una conversión a un espacio bidimensional haciendo uso del algoritmo t-SNE.

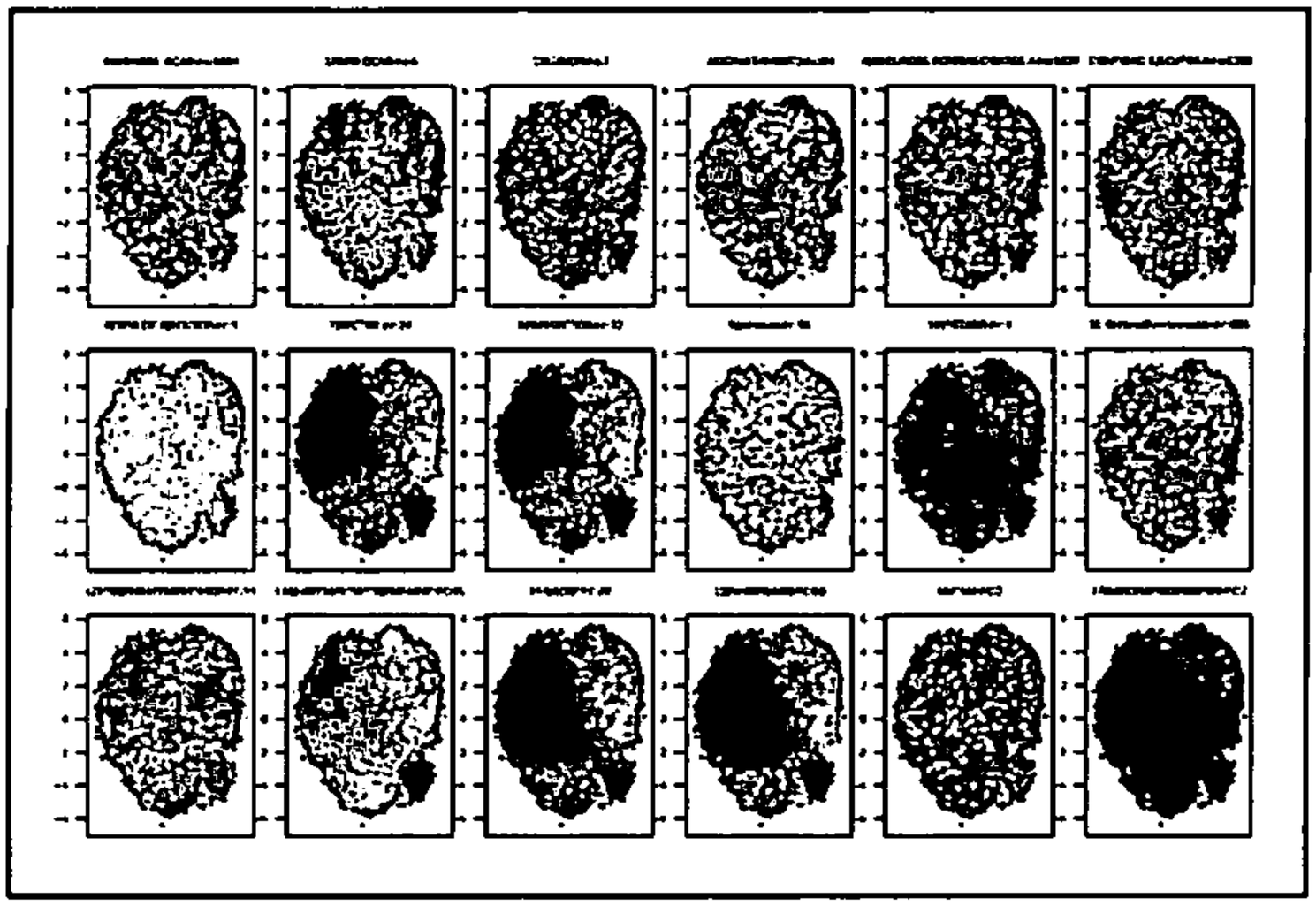


Figura 8. Visualización de los vectores en una representación 2D mediante el uso de t-SNE.

En la figura cada punto representa un proyecto y su color corresponde al valor de la variable categórica con la que se esté graficando.

Modelo clasificador

Para la construcción del modelo se hacen múltiples particiones del dataset, de tal forma que se divida en dos: un conjunto de entrenamiento con el 70% de los datos y un conjunto de prueba con el 30% de los datos.

Se realizaron pruebas con tres modelos: regresión logística, random forest y support vector machine (SVM). Se tomo el clasificador que en promedio entrega menor porcentaje de error en la clasificación del conjunto de prueba. El modelo final es un SVM con lo siguiente configuración:

- Kernel: RBF
- C = 1
- F1 score: 0,65
- MSE-training: 0,19
- MSE-test: 0,33

Conclusiones

- El análisis del texto complementa la información cuantitativa.
- Descubrir contexto a través de la vectorización.
- Evidenciar fallas "tempranas" en la formulación (correctivos anticipados).
- Priorización del seguimiento
    - Clasificar proyectos o entidades ejecutoras
    - Optimización de los recursos
- Clasificador adaptable a otras métricas de desempeño SGR

Anexo de Registros únicos

Tabla 1. Resumen información de registros encuesta y búsqueda web.

Entidad	Registro		Ori	RU?	Acceso	Tipo reg.	Matriz
DEPARTAMENTO ADMINISTRATIVO PARA LA PROSPERIDAD SOCIAL		LLAVE MAESTRA	E	Si	No	PN	No
Parques Nacionales Naturales de Colombia	RUNAP	Registro Único Nacional de Áreas Protegidas	E/B	Si	No	Otros	No
DIRECCIÓN NACIONAL DE DERECHO DE AUTOR		Registro Nacional de Obras	E	Si	No	Obras	No
SUPERINTENDENCIA DE LA ECONOMÍA SOLIDARIA	SICSES		E	No	No	Doc	No
FONDO ROTATORIO DE LA POLICÍA NACIONAL	BIOSTAR -		E	No	Registro	PN	Si
UNIDAD ADMINISTRATIVA ESPECIAL CONTADURÍA GENERAL DE LA NACIÓN	CUIN	Código Único Institucional	E	No	No	PJ	No
CONSEJO SUPERIOR DE LA JUDICATURA - DIRECCION EJECUTIVA ADMINISTRATIVA JUDICIAL	URNA	REGISTRO NACIONAL DE ABOGADOS	E	Si	No	PN	Si
DEPARTAMENTO ADMINISTRATIVO DE LA FUNCIÓN PÚBLICA	SUIT	Sistema Único de Información de Trámites	E	No	No Aplica	Trámites	No
DEPARTAMENTO ADMINISTRATIVO DE LA FUNCIÓN PÚBLICA	SIGEP	Sistema de Información y Gestión del Empleo Público	E	Si	Registro	PN	Si
MINISTERIO DEL INTERIOR Y DE JUSTICIA		Registro público de entidades religiosas no católicas	E	Si	No	PJ/PN	No
DIRECCION GENERAL DE LA POLICIA NACIONAL	RNMC	Registro Nacional de Medidas Correctivas	E	No	No	PN/Doc	No
MINISTERIO DE MINAS Y ENERGÍA	SICOM	Sistema de Información de Comercialización de Combustibles	E	No	Registro	Varios	No
UNIDAD ADMINISTRATIVA		Registro de Entidades	E	Si	No	PJ	No

ESPECIAL DE ORGANIZACIONES SOLIDARIAS		Acreditadas					
AGENCIA NACIONAL DE MINERIA		Registro único de Comercializadores de Minerales RUCOM	E	Si	Registro	PJ/PN	No
SUPERINTENDENCIA FINANCIERA DE COLOMBIA - SIMEV	RNVE	Registro Nacional de Valores y Emisores	E	Si	No	PJ	Si
SUPERINTENDENCIA FINANCIERA DE COLOMBIA - SIMEV	RNAMV	Registro Nacional de Agentes del Mercado de Valores	E	Si	No	PJ	Si
SUPERINTENDENCIA FINANCIERA DE COLOMBIA - SIMEV	RNPMV	Registro Nacional de Profesionales del Mercado de Valores	E	Si	No	PN	Si
SUPERINTENDENCIA FINANCIERA DE COLOMBIA	SIRI	Sistema Integrado de Registro de Información	E	No	No	Trámites	No
SUPERINTENDENCIA FINANCIERA DE COLOMBIA	SICOP	Sistema de Información de Colocaciones Ofertas Públicas	E	No	No	Trámites	No
REGISTRADURIA NACIONAL DEL ESTADO CIVIL - GESTION GENERAL	ANI	Archivo Nacional de Identificación	E	Si	No	PN	Si
REGISTRADURIA NACIONAL DEL ESTADO CIVIL - GESTION GENERAL	AFIS	Sistema Automatizado de Identificación Dactilar	E	No	No	Huellas	No
SUPERINTENDENCIA DEL SUBSIDIO FAMILIAR		CERTIFICADO DE EXISTENCIA Y REPRESENTACIÓN LEGAL	E	No	No	N/A	No
UNIDAD NACIONAL PARA LA GESTIÓN DEL RIESGO DE DESASTRES	RUD	Registro Único de Damnificados	E/B	Si	No	PN	Si
INSTITUTO NACIONAL PENITENCIARIO Y CARCELARIO - INPEC	SISIPEC	SISTEMA DE INFORMACIÓN PENITENCIARIO Y CARCELARIO	E	No	No	N/A	No
ARCHIVO GENERAL DE LA NACIÓN	RUSD	Registro Único de Series Documentales	E/B	Si	No	Doc	No
DEPARTAMENTO ADMINISTRATIVO DE CIENCIA, TECNOLOGÍA E INNOVACIÓN	CVLAC	Hojas de vida de investigadores	E	Si	Registro	PN	Si
DEPARTAMENTO ADMINISTRATIVO DE CIENCIA,	GRUPLAC	Información de grupos de investigación de	E	Si	Registro	Grupos	No

TECNOLOGÍA E INNOVACIÓN		CTI						
DEPARTAMENTO ADMINISTRATIVO DE CIENCIA, TECNOLOGÍA E INNOVACIÓN	INSTITULAC	Información de instituciones del SNCTI	E	Si	No	PJ	No	
DEPARTAMENTO ADMINISTRATIVO DE CIENCIA, TECNOLOGÍA E INNOVACIÓN	SIGP	Sistema Integrado de Gestión de Proyectos	E	No			No	
DEPARTAMENTO ADMINISTRATIVO DE CIENCIA, TECNOLOGÍA E INNOVACIÓN	Publindex	Sistema de Indexación y Homologación de Revistas Especializadas de CTI	E	Si	No	Revistas	No	
SUPERINTENDENCIA DE SOCIEDADES	RUNF	Registro Único Nacional de Factores	E/B	Si	No	PJ	No	
UNIDAD ADMINISTRATIVA ESPECIAL DIRECCIÓN DE IMPUESTOS Y ADUANAS NACIONALES	RUT	Registro único tributario	E/B	Si	Registro	PJ/PN	Si	
DIRECCIÓN DE ADMINISTRACIÓN DE FONDOS DE LA PROTECCIÓN SOCIAL	BDUA	Base de Datos Única de Afiliados	E	Si	No	PN	Si	
MINISTERIO DE EDUCACIÓN NACIONAL	SINEB	Sistema de Información Nacional de Educación Básica y Media	E	Si	No	Varios	Si	
MINISTERIO DE EDUCACIÓN NACIONAL	SNIES	Sistema Nacional de Información de la Educación Superior	E	Si	No	Varios	No	
INDUSTRIA MILITAR		Registro de Proveedores	E	Si	Registro	PJ/PN	No	
Unidad Administrativa Especial Migración Colombia		Registro de extranjeros	E	Si	No	PN	No	
UNIDAD ADMINISTRATIVA ESPECIAL DE GESTIÓN DE RESTITUCIÓN DE TIERRAS DESPOJADAS		Sistema de Registro de Tierras Despojadas y Abandonadas Forzosamente	E	N/A	N/A	N/A	No	
FIDUCIARIA COLOMBIANA DE		REGISTRO DE PROVEEDORES	E	Si	Registro	PJ/PN	Si	

COMERCIO EXTERIOR S.A.								
MINISTERIO DE TECNOLOGÍAS DE LA INFORMACIÓN Y LAS COMUNICACIONES	VUTIC	Ventanilla Única de Trámites	E	No	Registro	Trámites	No	
INSTITUTO NACIONAL DE MEDICINA LEGAL Y CIENCIAS FORENSES	RND	Registro Nacional de Desaparecidos y Cadáveres, Registro de Lesiones no Fatales	E	Si	No	PN	No	
MINISTERIO DE COMERCIO, INDUSTRIA Y TURISMO		Registro de Usuario de la VUCE	E	Si	No	PJ/PN	No	
MINISTERIO DE COMERCIO, INDUSTRIA Y TURISMO	RUNIC	Registro Único Nacional de Importadores y Comercializadores Autorizados	B	Si	No	N/A	No	
FUERZA AÉREA COLOMBIANA	SIATH		E	N/A	N/A	N/A	No	
BANCO DE COMERCIO EXTERIOR DE COLOMBIA S.A.		Índice de Información Clasificada y Reservada	E	N/A	N/A	N/A	No	
BANCO DE COMERCIO EXTERIOR DE COLOMBIA S.A.		Índice de Información Clasificada y Reservada	E	N/A	N/A	N/A	No	
SUPERINTENDENCIA DE SERVICIOS PÚBLICOS DOMICILIARIOS	RUPS	REGISTRO UNICO DE PRESTADORES DE SERVICIOS PUBLICOS	E	Si	No	PN	Si	
INSTITUTO COLOMBIANO AGROPECUARIO		Registro Nacional de Predios Pecuarios	E	Si	No	Otro	No	
INSTITUTO DE HIDROLOGIA, METEOROLOGIA Y ESTUDIOS AMBIENTALES- IDEAM	RUA	Registro único ambiental	E/B	Si	Registro	Otro	Si	
UNIDAD ADMINISTRATIVA ESPECIAL BIBLIOTECA NACIONAL		REGISTRO ÚNICO DE BIBLIOTECAS PUBLICAS	E	Si	No	N/A	No	
Instituto Nacional de Salud	sivigila	Sistema Nacional de Vigilancia en Salud Pública	E	No	No	Otro	No	
Instituto Nacional de Salud	sivicap	Sistema de Información de la Vigilancia de la Calidad del Agua	E	No	No	Otro	No	

		para Consumo Humano						
Instituto Nacional de Salud	Reddata INS	Información donantes de órganos	E	N/A	N/A	N/A		No
MINISTERIO DE CULTURA		Registro como productor permanente u ocasional de espectáculos públicos de las artes escénicas	E	Si	Registro	PJ/PN		No
MINISTERIO DE CULTURA		Autorización de Operadores de Boletería en Línea	E	No	N/A	PJ/PN		No
UNIDAD ADMINISTRATIVA ESPECIAL DE GESTIÓN PENSIONAL Y CONTRIBUCIONES PARAFISCALES DE LA PROTECCIÓN SOCIAL	RUA	Registro único de aportantes	E	Si	No	PN		No
MINISTERIO DEL TRABAJO		40.00 empleos	E	No	No			No
FASECOLDA	RUS	Registro Único de Seguros	B	Si	No	PJ		No
Ministerio de Salud y Protección Social	RETHUS	Registro Único de Talento Humano en Salud	B	Si	Registro	PN		Si
MINISTERIO DE TRABAJO	RUTEC	Registro Único de Trabajadores Extranjeros en Colombia (En construcción)	B	Si	No	PN		No
MINISTERIO DE EDUCACIÓN NACIONAL		Registro Único de Peticiones	B	Si	No	Trámites		No
CONFECAMARAS	RUNEOL	Registro Único Nacional de Entidades Operadoras de Libranza	B	Si	Registro	PJ		Si
Ministerio de Salud y Protección Social	RUAF	Registro Único de Afiliados (Tramite, afiliación, identificación)	B	Si	No	PN		Si
Cancillería	RegRetor	Registro Único de Retornados	B	Si	Registro	PN		Si
Unidad para las víctimas - RNI	RUV	Registro Único de Víctimas	B	Si	No	PN		Si
MINISTERIO DE TRANSPORTE	RUNT	Registro Único Nacional de Tránsito	B	Si	Registro	PJ/PN/Vehículos		Si
Cámaras de Comercio	RUP	Registro Único de Proponentes	B	Si	Registro	PJ/PN		Si

		(Info. Ubicación)					
MINISTERIO DE TRABAJO	RUI	Registró Único de Intermediarios	B	Si	Registro	PJ/PN	Si
Agencia Nacional Minera	RUCOM	Registro Único de Comercializadores de Minerales	B	Si	Registro	PJ/PN	Si
CONFECAMARAS	RUES	Registro Único Empresarial y Social, Cámaras de Comercio	B	Si	No	PJ/PN	Si
CONFECAMARAS	RNT	Registro Nacional de Turismo	B	Si	No	PJ/PN	Si
INSTITUTO DE INVESTIGACIÓN ALEXANDER VON HUMBOLDT	RNC	Registro Único Nacional de Colecciones Biológicas	B	Si	No	Otro	Si

Tabla 2. Ranking de campos de acuerdo al número de repeticiones en el conjunto y al IRC respectivo

CAMPO	Rep.	σ_1		CAMPO	Rep.	σ_1
No. De documento / identificación	24	0,88		Cargo	2	0,04
Primer nombre (nombres)	24	0,88		Parentesco	1	0,00
Tipo de documento	21	0,77		Libreta / Tarjeta profesional	1	0,00
Primer apellido (apellidos)	21	0,77		Estado tarjeta prof.	1	0,00
Segundo nombre (Op)	17	0,62		Motivo	1	0,00
Segundo apellido (Op)	17	0,62		Libreta Militar	1	0,00
Ubicación / Dirección	17	0,62		Afro	1	0,00
Correo electrónico	16	0,58		Familia No.	1	0,00
Departamento / Ciudad/municipio	16	0,58		Estado salud	1	0,00
Datos de contacto (Teléfonos, Dirección)	15	0,54		Otra nacionalidad	1	0,00
NIT	13	0,46		Pag. Web	1	0,00
Notificaciones (Dir, tel, email)	12	0,42		Responsabilidades	1	0,00
Razón Social	10	0,35		Lugar Declaración	1	0,00
Nombre Comercial Sigla / Nombre establecimiento	8	0,27		Fecha Declaración	1	0,00
País	7	0,23		Declaración Caso Especial	1	0,00
Genero / Sexo	6	0,19		Se considera víctima (Tipo acto)	1	0,00
Fecha de nacimiento	6	0,19		Tipo de tramite	1	0,00
Fecha de solicitud / elaboración / inscripción	6	0,19		Tipo de afiliación	1	0,00
Tipo persona / Tipo	5	0,15		Tipo de cotizante	1	0,00

Titular						
Pertenencia Étnica / Origen étnico	5	0,15		Depto. / Ciudad	1	0,00
Ciudad/municipio nacimiento	5	0,15		Nivel Sisben	1	0,00
Corregimiento / Vereda / Sector / Barrio	5	0,15		Población especial	1	0,00
Actividad ppal / Actividad económica	5	0,15		Adm. ARL	1	0,00
País de nacimiento	4	0,12		Adm. Pensiones	1	0,00
Entidad que atiende / Cámara de Comercio	4	0,12		IBC	1	0,00
Nacionalidad	3	0,08		Matricula / Renovación / Inscripción	1	0,00
Lugar de expedición	3	0,08		Datos licencia de conducción (No, exp., fecha exp., estado, restricción)	1	0,00
Fecha de expedición	3	0,08		Multas e infracciones	1	0,00
Régimen	3	0,08		Certificado medico	1	0,00
Tipo de organización / Tipo establecimiento	3	0,08		Certificado de aptitud	1	0,00
Titulo pregrado	3	0,08		Info solicitudes	1	0,00
Universidad pregrado	3	0,08		Tipo intermediario	1	0,00
Lugar pre-grado	3	0,08		Minerales	1	0,00
Fecha de pre-grado	3	0,08		Ed. básica y media	1	0,00
Lugar postgrado	3	0,08		Titulo obtenido	1	0,00
Fecha de postgrado	3	0,08		Fecha grado	1	0,00
Titulo postgrado	3	0,08		Organización Jurídica	1	0,00
Universidad postgrado	3	0,08		Operador / Administrador: Libranza	1	0,00
Estado civil	2	0,04		Entidad	1	0,00
Datos de contacto (Teléfonos, Dirección, email) - Alternos	2	0,04		Estado RNPMV	1	0,00
Tipo de afiliado	2	0,04		Modalidad	1	0,00
Discapacidad	2	0,04		Especialidad	1	0,00
Estado	2	0,04		Vigencia desde	1	0,00
Numero de inscripción / Num. Matricula	2	0,04		Vigencia hasta	1	0,00
Actividad secundaria	2	0,04		Fecha inscripción	1	0,00
Otras actividades	2	0,04		Cod. institución	1	0,00
Idiomas	2	0,04		Ppal o seccional	1	0,00
Experiencia laboral	2	0,04		Naturaleza jurídica	1	0,00
Código tipo de entidad RNVEI	2	0,04		Sector	1	0,00
Código RNVEI	2	0,04		Carácter académico	1	0,00

Calidad Inscrito	2	0,04		Norma de creación	1	0,00
Estado RNVEI	2	0,04		Fecha de la norma	1	0,00
Situación legal	2	0,04		Acreditada alta calidad	1	0,00
Dirección Web	2	0,04		Fecha acreditación	1	0,00
Fax	2	0,04		Resolución acreditación	1	0,00
No. miembros junta directiva	2	0,04		Vigencia acreditación	1	0,00
Representante legal	2	0,04		Cobertura	1	0,00
Entidad vigilante	2	0,04				

SGR
Archivo avances OCAD

BPIN	Llave (Identificador único)	Descripción
NOMBRE OCAD	Categórica	(Depto. – Mun.)
TIPO OCAD	Categórica	Ciencia y Tecnología, CorMagdalena, Corporaciones, Departamental, Municipal, Regional.
REGION	Categórica	Caribe, Centro Oriente, Centro Sur, CorMagdalena, Del Llano, Eje Cafetero, Pacífico
DEPARTAMENTO	Categórica	
RECURSOS PERTENECIENTES	Categórica	(Depto + Muni)
CODIGO DANE ENTE EJECUTOR	Numérico	
ENTIDAD EJECUTORA	Categórica	Depto., Mun, Corporación, Univ.
TIPO EJECUTOR	Categórica	Depto., Mun., CAR, MinasEnergia, Otros, SinBPIN, SinEjecutorDesignado
SECTOR	Categórica (21)	
SUBSECTOR	(73)	
NOMBRE DEL PROYECTO	Texto	
AVANCE FÍSICO	Numérico	
AVANCE FINANCIERO	Numérico	
ESTADO	Categórico (10)	
FECHA DE APROBACION	Date	
FTE SGR DIRECTAS	Numérico	
FTE SGR FCR_40	Numérico	
FTE SGR FCR_60	Numérico	
FTE SGR FDR	Numérico	
FTE SGR FCTEI	Numérico	

FTE SGR CORMAGDALENA	Numérico
FTE SGR INCENTIVO A LA PRODUCCION	Numérico
TOTAL VALOR SGR	Numérico
VALOR NACION	Numérico
VALOR OTROS	Numérico
VALOR TOTAL PROYECTO	Numérico

8. Secretaría de transparencia

Participantes: José Zea. Bajo la coordinación de Mario Morales

Comentarios generales

1. Construir una base de datos relacional que conecte las denuncias con las remisiones
2. Construir etiquetes de valor y de variables para todas las variables
3. Utilizar un número predeterminado de categorías para cada variable, minimizar las variables de texto
4. Utilizar la División Política Administrativa para

Comentarios específicos

1. Colocar nombres a las variables más explícito:

   - ✓ "TIPO RADICADO" por ID_ORIGEN_DENUN
   - ✓ FECHA por FECHA_DENUNCIA
   - ✓ ENTIDAD por ENTIDAD_DENUNCIA
   - ✓ "DEPTO ORIGEN DE LA DENUNCIA" por DEPTO_ORIGEN_DENUNCIA
   - ✓ MUNICIPIO por MUNICIPIO_ORIGEN_DENUNCIA
   - ✓ Presunto acto de Corrupción o Irregularidad por IRREGULARIDAD
   - ✓ "TIPO DE FORMA DE CONTRATACIÓN para denuncias precontractuales (Licitación, Directa, NA. Etc.) por TIPO_CONTRATACIÓN
   - ✓ ENTIDAD MENCIONADA EN LA DENUNCIA por ENTIDAD_MENCIONADA_DENUNCIA
   - ✓ ORDEN DE LA ENTIDAD por ORDEN_ENTIDAD_MENCIONADA_DENUNCIA
   - ✓ SECTOR por SECTOR_ENTIDAD_MENCIONADA_DENUNCIA

2. Hay 42 faltantes en las denuncias

```
table(duplicated(denuncias$ID_ORIGEN_DENUNCIA))
```

```
FALSE  TRUE
 5755    42
```

3. Para la variable de las denuncias precontractuales hay 2420 faltantes, sería deseable colocar una categoría No contractual

```
> table(denuncias$TIPO_CONTRATACIÓN, useNA = "always")

    CONTRACTUAL-OTROS CONTRATOS          CONTRACTUAL -LICITACIÓN PRECONTRACTUAL- OTROS CONTRATOS
                            945                              151                             1522
     PRECONTRACTUAL -LICITACIÓN                             <NA>
                            717                             2420
```

4. Cargar la División política Administrativa para los departamentos y municipio de donde provienen las denuncias. Lo anterior es muy importante ya que arreglar la ortografía de los municipios es un proceso muy engorroso y la recategorización de los municipios se hace muy complicada.

5. Almacenar la fecha en un formato tipo fecha para evitar errores.

6. El *orden* de la entidad mencionada y la variable Presunto *acto de Corrupción o Irregularidad* se debe estandarizar (mayúsculas o minúsculas)

7. En la variable orden de la entidad se tiene que en la categoría turismo sólo tiene un caso y existe una categoría otros bastante extensa, revisar si es justificable abrir esta categoría.

8. Re-categorizar las variables con categorías no tan largas, por ejemplo, en la variable Presunto acto de Corrupción las categorías son muy largas.

	Var1	Freq
4	DEFICIENCIAS EN LA PLANEACIÓN (ESPECÍFICAS)	1709
20	TRÁFICO DE INFLUENCIAS (GENERAL) ART. 431 Y 4...	1043
13	MALA ADMINISTRACIÓN Y/O MALVERSACIÓN DE RE...	771
1	ABUSO DE AUTORIDAD ART. 416 Y 417 DEL C.P.	588
2	CELEBRACIÓN DE CONTRATOS SIN EL LLENO DE R...	435
12	MALA ADMINISTRACIÓN Y MALVERSACIÓN DE FON...	207
10	INCUMPLIMIENTO LEY DE TRANSPARENCIA	175

9. Sería interesante crear una categoría de tipo de ENTIDAD_DENUNCIANTE, (para que se puedan identificar rápidamente que denuncias son ciudades, de alcaldes, de veedurías, del concejo, del congreso, comunales, empresas, etc.) ya que con el nombre de la entidad denunciante es algunas veces complicado identificar el tipo de entidad que realizó la denuncia para efectos de análisis y modelamiento.

	ENTIDAD_DENUNCIANTE	Freq	Freq%	Freq%Acum
1	CIUDADANO	3510	38.48998831	38.48997
2	DE	269	2.94826836	41.43824
3	SAS	207	2.26874178	43.68698
4	VEEDOR	192	2.10454028	45.79152
5		160	1.85225778	47.64378
6	UTDA	125	1.37001315	49.01379
7	CONSORCIO	101	1.10687063	50.12066
8	Y	99	1.08505042	51.20571
9	C	73	0.80006768	52.00578
10	CONSTRUCCIONES	52	0.56997547	52.57576
11	LA	52	0.56997547	53.14574
12	COLOMBIA	50	0.54808526	53.69383

Showing 1 to 13 of 1,342 entries

10. En la hoja de cálculo de SOBRES ENVIADOS hay datos faltantes en las columnas pares:

	FECHA - ENVIO	Respuesta	DESTINATARIO	DEPENDENCIA	CIUDAD	CLASE	FOLIOS	FIRMO
1	2014-01-02	OFI13-00150132	RENE RODRIGUEZ	CIUDADANO	BARRANCABERMEJA	DENUNCIAS	1	ALIRK
2	*NA*	*NA*	*NA*	*NA*	*NA*	*NA*	*NA*	*NA*
3	2014-01-03	OFI14-00000227	PEDRO RODRIGUEZ	CIUDADANO	CARTAGENA	DENUNCIAS	1	ALIRK
4	*NA*	*NA*	*NA*	*NA*	*NA*	*NA*	*NA*	*NA*
5	2014-01-03	OFI14-00000261	JORGE ARRIETA	CIUDADANO	COROZAL	DENUNCIAS	1	ALIRK
6	*NA*	*NA*	*NA*	*NA*	*NA*	*NA*	*NA*	*NA*

11. Hay faltantes en la variable identificadora de la base de datos de los sobres (número de respuesta).
12. Renombrar de forma adecuada los nombres de la base de datos de sobres
13. Arreglar la Fecha en formato tipo fecha
14. En lo sobres, arreglar las clases para que existan unas únicas categorías:

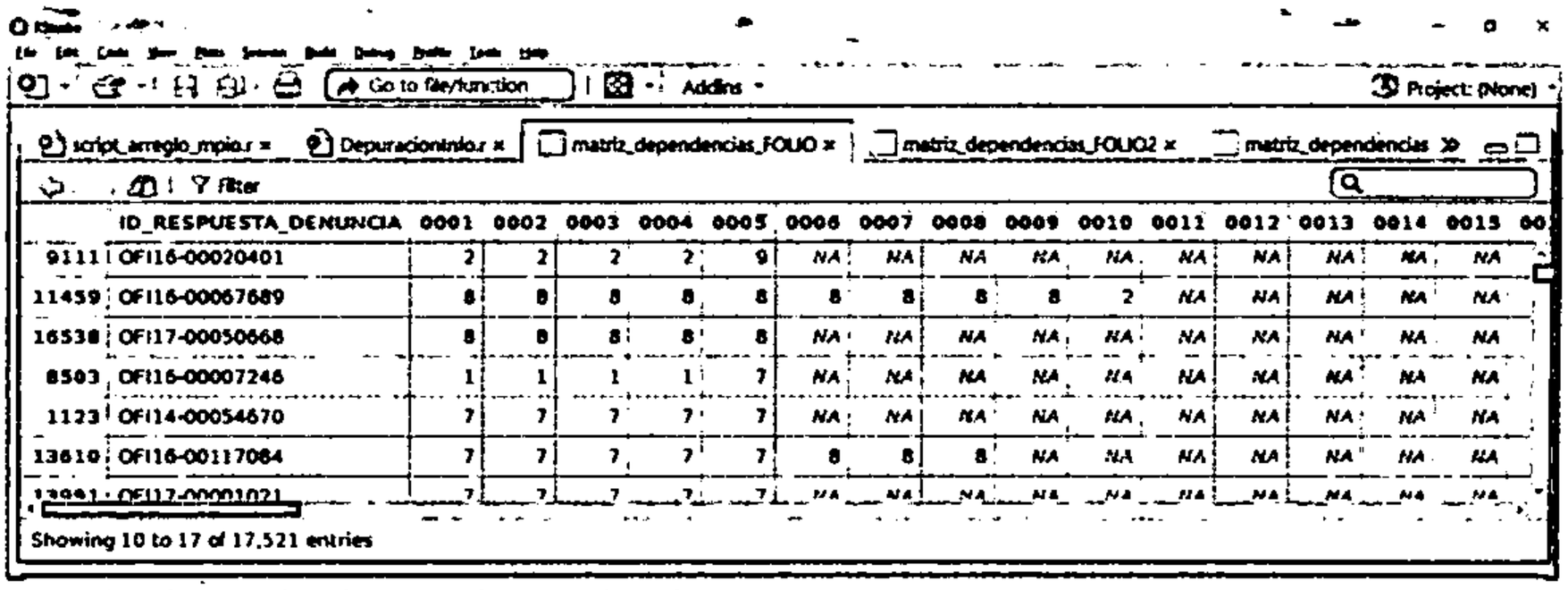
Por ejemplo, ADMINISTRATIVO tiene las siguientes grafías: ADMIINISTRATIVO, ADMINISTRATIVA, Administrativo, ADMINISTRATIVO, ADMINISTRSATIVO, ADMINSTRATIVO, ADMNIISTRATIVO.
Para DENUNCIAS se encuentran las siguientes gráficas: DENUNCIAS, DENUNCIAS, DENUNCIAS, DENUNCIAS, DENUNCIAS, DENUNCIAS, DENUNCIAS, DENUNCIAS, DENUNCIAS, DENUNCIAS, DENUNCIAS, DENUNCIAS, DENUNCIAS.

15. Las denuncias son remitidos a múltiples fuentes:

	ID_RESPUESTA_DENUNCIA	0001	0002	0003
791	OFI15-00100851	GOBERNADOR	FUERZA AEREA DE COLOMBIA	AERONAL
164	OFI16-00063670	GERENTE FONDO DE ADAPTACIÓN	SECRETARIO DE OBRAS PÚBLICAS ALCALDIA DE YO...	AGENTE I
133	OFI15-00037355	AGRADECIMIENTO LEY159 DE 2014	AGRADECIMIENTO LEY159 DE 2014	AGRADEC
146	OFI15-00084041	ALCALDE	ALCALDE	ALCALDE
122	OFI16-00037629	ALCALDE DE GUAMAL	ALCALDE DE GRANADA	ALCALDE
161	OFI17-00006897	ALCALDE LOCAL-LOCALIDAD INDUSTRIAL Y DE LA B...	PERSONERO DISTRITAL DE CARTAGENA	ALCALDE

Showing 1 to 7 of 17,521 entries

16. Existen algunos casos en que Los folios varían en tamaño de remisión a remisión

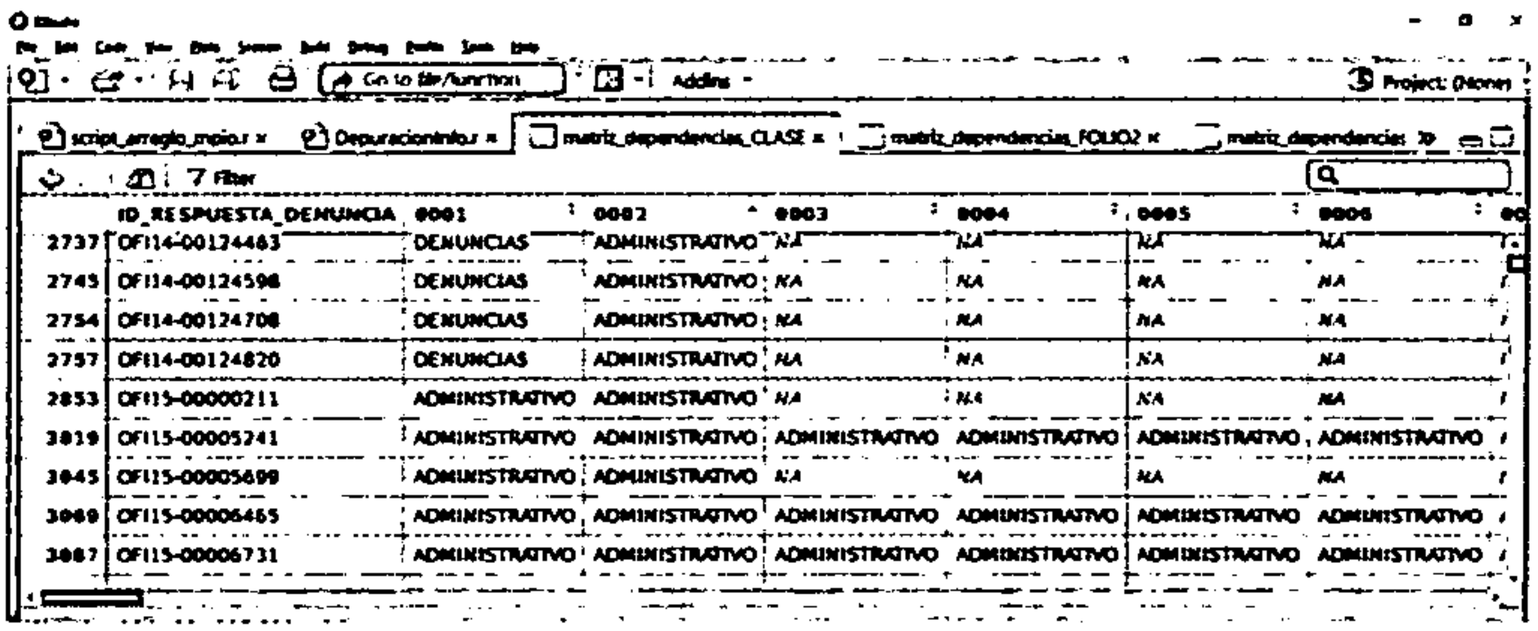


	ID_RESPUESTA_DENUNCIA	0001	0002	0003	0004	0005	0006	0007	0008	0009	0010	0011	0012	0013	0014	0015	00
9111	OFI16-00020401	2	2	2	2	9	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	
11459	OFI16-00067689	8	8	8	8	8	8	8	8	8	2	NA	NA	NA	NA	NA	
16538	OFI17-00050668	8	8	8	8	8	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	
8503	OFI16-00007246	1	1	1	1	7	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	
1123	OFI14-00054670	7	7	7	7	7	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	
13610	OFI16-00117084	7	7	7	7	7	8	8	8	NA	NA	NA	NA	NA	NA	NA	
13981	OFI17-00001021	7	7	7	7	7	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	

Showing 10 to 17 of 17,521 entries

Se trabajó con el máximo número de páginas

17. Existen algunas clases en que la clase del sobre varía en los remitentes:



	ID_RESPUESTA_DENUNCIA	0001	0002	0003	0004	0005	0006	00
2737	OFI14-00124483	DENUNCIAS	ADMINISTRATIVO	NA	NA	NA	NA	
2745	OFI14-00124598	DENUNCIAS	ADMINISTRATIVO	NA	NA	NA	NA	
2754	OFI14-00124708	DENUNCIAS	ADMINISTRATIVO	NA	NA	NA	NA	
2757	OFI14-00124820	DENUNCIAS	ADMINISTRATIVO	NA	NA	NA	NA	
2853	OFI15-00000211	ADMINISTRATIVO	ADMINISTRATIVO	NA	NA	NA	NA	
3819	OFI15-00005241	ADMINISTRATIVO	ADMINISTRATIVO	ADMINISTRATIVO	ADMINISTRATIVO	ADMINISTRATIVO	ADMINISTRATIVO	
3845	OFI15-00005699	ADMINISTRATIVO	ADMINISTRATIVO	NA	NA	NA	NA	
3969	OFI15-00006465	ADMINISTRATIVO	ADMINISTRATIVO	ADMINISTRATIVO	ADMINISTRATIVO	ADMINISTRATIVO	ADMINISTRATIVO	
3987	OFI15-00006731	ADMINISTRATIVO	ADMINISTRATIVO	ADMINISTRATIVO	ADMINISTRATIVO	ADMINISTRATIVO	ADMINISTRATIVO	

18. En el archivo que relaciona la denuncias con los envíos hay duplicados por el identificados de los sobres (RESPUESTA):

```
> table(duplicated(correlativa$ID_RESPUESTA_DENUNCIA))

FALSE  TRUE
35671  5923
```

19. De las 5755 denuncias analizadas, cruzaron el 77 % de las denuncias correspondientes al 77% de las denuncias:

9. Red de usuarios empresariales en Colombia

Participantes: Daniel Rojas y Juan David Martínez. Bajo la coordinación de Santiago Matallana y Mario Morales.

Justificación

El periódico la República realiza una encuesta a 800 empresarios de todo el país calificando de 1 a 5 al gabinete presidencial, dicha encuesta no cuenta con una metodología robusta puesto que son preguntas de percepción frente a ministros y altor cargos gerenciales del Gobierno.

Luis Fernando Mejía solicitó al equipo de comunicaciones junto con la DDE en desarrollar una estrategia que logre subir su puntaje dado que en la encuesta pasada sacó 2.5 sobre 5 (el puntaje más alto lo sacó el Presidente Juan Manuel Santos con una calificación de 3.3 aproximadamente) por lo que la DDE y Comunicaciones han estado trabajando de la mano en generar dicha estrategia y a puertas de tenerla ya lista la DDE identificó la posibilidad de realizar un análisis en redes sociales sobre los usuarios que mayor participación e influencia tienen dentro del sector empresarial y de esa manera involucrar un componente de marketing dirigido a aquellos que tengan un número considerable de seguidores y puedan mejorar la percepción del Director mediante mensajes de invitación a conocer lo que el DNP está haciendo y lo que Luis Fernando Mejía hace por el sector empresarial.

Metodología

Con el fin de conocer de qué manera está compuesta la red de usuarios que hablan sobre temas empresariales en el país, junto a la Dirección de Desarrollo Empresarial (DDE) se creó un set de palabras relacionadas al sector empresarial y junto a la Dirección de Comunicaciones se capturaron los usuarios en Twitter de los principales representantes del sector empresarial para cada industria.

Se procedió a hacer la captura de tweets el día 29 de agosto a las 2:00 pm y se culminó la captura el día 30 de agosto a las 2:00 p.m. teniendo un periodo muestra de 24 horas en donde se obtuvieron 92.226 tweets y 63.059 usuarios para analizar.

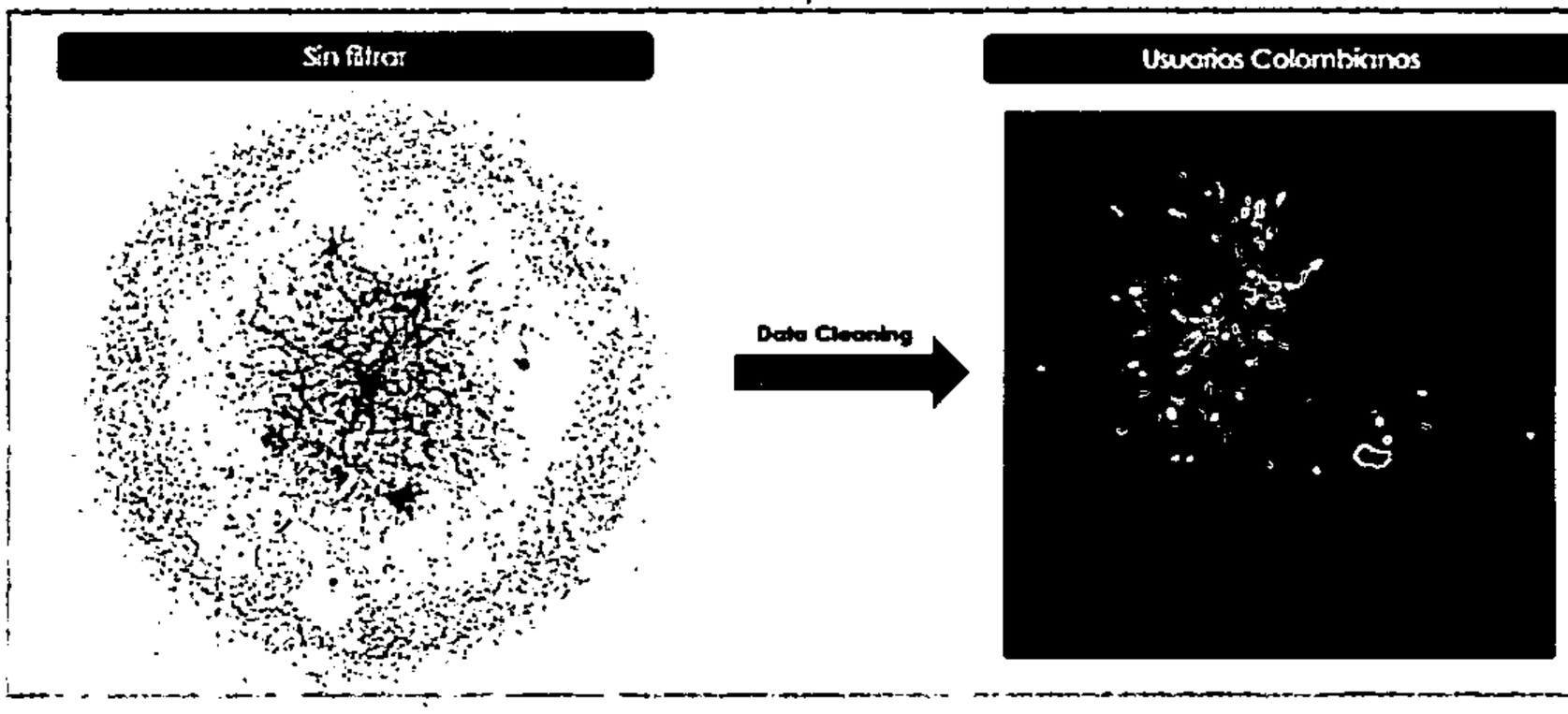
Dentro de las métricas utilizadas para valorar la centralidad de los usuarios en la red, se tienen:

- Distribución de grados tanto de entrada / salida:
  Se calcula el diámetro de los nodos con base en el número de conexiones entrantes que este tenga

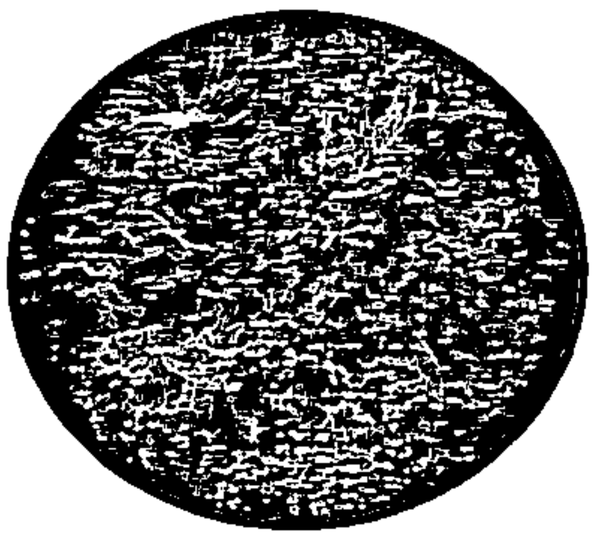
- Modularidad:
  Aleatoriamente se selecciona un grupo de nodos para cuantificar la densidad de sus conexiones y contrastar dicha densidad con la de toda la red, si la densidad es mayor que la de toda la red se considera una comunidad
- Algoritmo de PageRank:
  Con base en el número de seguidores que tienen los usuarios que trinan, se asigna una calificación al usuario citado para determinar el diámetro de su nodo

Conclusiones

La red inicial contaba con un gran número de usuarios y tweets no relacionados con el país o involucrando otro tipo de temáticas ajenas a Colombia, por lo que se tuvo que realizar una limpieza profunda de cada perfil y limitar el número de usuarios a analizar solo a aquellos que mencionaran ciudades o municipios de Colombia o tuviesen sus perfiles relacionados con alguna ubicación dentro del país por lo que luego de limpiar los datos se tuvieron 1.971 usuarios en total para Colombia.



Dentro de los resultados más resaltantes capturados con page rank y modularidad se obtuvieron los siguientes rankings:



Top de Usuarios con mayor popularidad en la red (PageRank)

Puesto	Usuario
1	@guillermobotero
2	@caroco
3	@blurodioco
4	@uinvestigativa
5	@unadavid
6	@nunezluis
7	@brucemacmaster
8	@zuluagajorge
9	@jrodriguezt
10	@universidaduptc

Comunidades Encontradas		
Porcentaje de la red	**Usuarios Representativos**	**Color**
20.14%	@guillermobotero, @blueradioco, @ulnvestigativa, @wradiocolombia	
16.74%	@brucemacmaster, @andi_Colombia, @sicsuper, @procolombiaco	
14.41%	@mincomercioco, @mciacouture, @confecamaras, @fonturcol	
13.85%	@zuluagajorge, @nunezluis, @corpoica, @mantillaignacio	
12.38%	@lunadavid, @crozo, @ministeriotic, @jrodriguezk	
5.68%	@anhcolombia, @marcela_amaya6, @mintrabajocol	
5.12%	@fedecafeteros, @masty_Sanchez, @ltzramirezc, @comitecafestder	
4.87%	@infopresidencia, @aurelioiragorri, @herman_amayat, @finagro	
3.25%	@alvarouribevel, @valorainversion, @camilosilva, @cedemocratico	
1.47%	@bogota, @carloscarrillos	
1.32%	@redadelco, @revistagrama, @galvalletenzano, @ccvalleduper	
0.76%	@caro_apoterw, @colbrinjton	

Por lo cual se justificó una estrategia directa a los usuarios con mayor afinidad en la red a Luis Fernando Mejía y una búsqueda constate de retweets por parte de estos mismos. Adicionalmente se recomendó identificar a los usuarios que fueron entrevistados en la última versión de la encuesta y seguirlos en las principales redes sociales buscando que ellos sigan al Director General y por consecuente a la cuenta oficial del DNP.