Análisis de indicadores TerriData – Fase 1

Unidad de Científicos de Datos

2018

ANÁLISIS DE INDICADORES TERRIDATA – FASE I

Entidad
Departamento Nacional de Planeación.

- Dirección de Desarrollo Digital.
- Dirección de Descentralización y Desarrollo Regional.

Sector	Lenguaje	Fuente de datos
Planeación.	R.	Terridata.

Presentación
Terridata es un proyecto multipropósito motivado por la buena disponibilidad de información a nivel municipal y de alcance nacional, contenida en una estructura de datos panel que centraliza información diversa y completa sobre las principales estadísticas asociadas a la situación de las entidades territoriales. Este proyecto busca capitalizar la oportunidad de establecer comparaciones y generar analítica a partir de información de diferentes categorías y fuentes. Se propusieron un conjunto de metodologías para explotar desde varias perspectivas la información estructurada y no estructurada contenida en Terridata, buscando pronosticar las variables, establecer relaciones causales entre niveles de inversión y desarrollo municipal y aplicar técnicas de minería de texto sobre los Planes de Desarrollo Territorial.

Terridata is a multipurpose project motivated by the availability of rich Municipality-level information gathered in a panel data structure that centralizes diverse and complete information relating to the status of all territorial entities. This project aims to exploit the opportunity to make comparisons and gain analytical insights from data belonging to different categories and sources. A set of methodologies are proposed to exploit the data contained in Terridata, seeking to predict variable series, to find causal relationships between investment levels and municipal development, and to find patterns in the goals set forth in local "Territorial Development Plans".

Objetivo general
Definir metodologías automáticas de análisis de la información contenida en la base de datos de Terridata.

Objetivos específicos

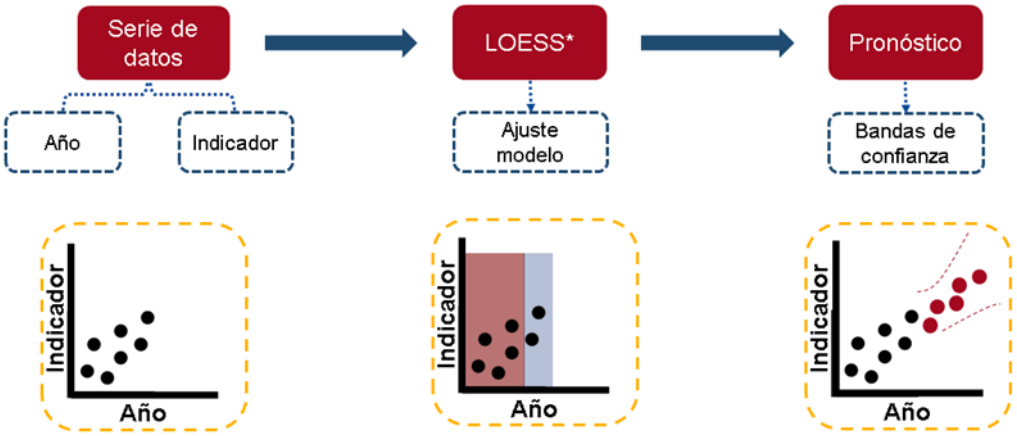
- Desarrollar una herramienta para pronosticar las variables de Terridata.
- Proponer una metodología estadística para estimar el impacto de las principales inversiones sobre los indicadores de desarrollo municipal.
- Estandarizar las metas de los Planes de Desarrollo Territorial a partir de la determinación de una serie de metas tipo por sector.

Metodología
Pronóstico de Variables
Terridata contiene más de 350 variables pertenecientes a 13 sectores, provenientes de entidades gubernamentales y no gubernamentales. Para cualquier entidad territorial, es posible definir una serie de tiempo para cualquier indicador, considerando las sucesivas mediciones en estricto orden temporal. El problema de pronóstico consiste en elaborar una predicción para la variable aleatoria *h* periodos adelante, donde *h* es el horizonte de pronóstico. En este proyecto se exploró el uso de la regresión polinomial local para elaborar pronósticos óptimos, respecto a un conjunto de información univariado y a diferentes horizontes de pronóstico.

En primer lugar, se define la serie de tiempo de interés, constituida por el conjunto de realizaciones de una variable para una entidad territorial específica y durante un periodo de tiempo definido, es decir:

$$\{y^m{}_t : t = 1, 2, \dots T\}$$

donde y_t^m es valor en el momento *t* (entre *T* posibles) de la variable y para la *m*-ésima entidad territorial. Metodológicamente, se asume una perspectiva de pronósticos con un enfoque de estadística aplicada. Por lo tanto, se considera que las observaciones de la variable aleatoria constituyen ejemplos o instancias de una estructura probabilista subyacente, cuyo comportamiento queremos modelar a partir de una representación matemática. En este caso, la representación consiste en un modelo de regresión polinomial local. Asumiendo que el modelo estimado es correcto, se pueden elaborar pronósticos de densidad y pronósticos puntuales del proceso de serie de tiempo subyacente.



Seleccionada una serie de tiempo de interés, se ajusta el modelo de regresión polinomial local, el cual consiste en una especificación paramétrica para la función de regresión, pero que otorga mayor ponderación a las observaciones más próximas al punto de referencia. Por lo tanto, la estimación consiste en determinar el modelo polinomial que mejor aproxima los datos en una vecindad arbitrariamente pequeña alrededor de cualquier momento de tiempo. La técnica de estimación es mínimos cuadrados ponderados.

En el pronóstico de la variable para el siguiente periodo, $y^m{}_{t+1}$, se otorga mayor ponderación a las mediciones más recientes de la variable aleatoria. Por lo tanto, el pronóstico está basado en la extrapolación del comportamiento local y no en el comportamiento global de toda la serie de tiempo.

Estimación del Impacto de las Inversiones

Territdata contiene información de los diferentes montos de inversión asignados a los municipios por las principales fuentes que constituyen el Sistema Nacional de Transferencias, en particular el Presupuesto General de la Nación (PGN) y el Sistema General de Regalías (SGR). Estas inversiones tienen una asignación sectorial específica (Educación, Salud, Saneamiento de Agua Potable, etc.). Desde un punto de vista económico, social y administrativo es relevante indagar por la eficiencia de las inversiones y el impacto causal que las mismas representan para la evolución de los indicadores en las cuáles debe reflejarse el Desarrollo Municipal. En este proyecto se implementó una metodología de Evaluación de Impacto basada en el emparejamiento (*Matching*), la cual permite establecer relaciones aproximadamente causales entre los niveles de inversión en las entidades territoriales y el Desarrollo Municipal observado a través de un conjunto de indicadores.

La asignación de diferentes montos de inversión a las entidades territoriales se inscribe dentro de lo que podríamos denominar una situación “no experimental”, debido a que la cantidad que se invierte en los municipios es no aleatoria *(non-random assignment)* y por el contrario responde a una lógica institucional implícita. Por ejemplo, el monto de las inversiones puede estar condicionado por la capacidad que tienen los municipios de señalizar eficiencia administrativa. En ese sentido, los municipios que tienen niveles

relativamente bajos de gasto pueden ser sistemáticamente diferentes y no constituir un contrafactual válido de los municipios que presentan niveles de inversión relativamente altos, lo cual dificulta la estimación del efecto causal del nivel de inversión sobre el desarrollo municipal. Esto se conoce en la literatura como el problema fundamental de la Evaluación del Impacto y para solventarlo existe un conjunto de metodologías de análisis, entre las cuales contamos el emparejamiento basado en observables (*Matching*).

Teniendo en cuenta la gran cantidad de información contenida en Terridata, el primer paso es una preselección de variables pertinentes a la operación estadística. Para ello, se determinó un conjunto reducido de indicadores que fuesen representativos del Desarrollo Municipal, las cuáles fueron seleccionadas de las categorías de Alcantarillado, Educación y Salud. Por otra parte, se consideró un conjunto de rubros de inversión que deberían afectar (en el sentido de estricta causalidad) estos indicadores, resultando en la conformación de una matriz de efectos teóricos que determinan los casos o instancias interesantes de evaluación de impacto. Nótese que, en un contexto con D indicadores de desarrollo municipal e I rubros de inversión, tenemos C=D*I casos (instancias) de evaluación de impacto, de los cuáles solo un subconjunto tiene relevancia teórica.

Indicador / Inversión	Educación	Salud	Agua Potable	Propósito general	Alimentación escolar	Ribereños	Resguardos indígenas
Cobertura de acueducto (Censo)			■			■	
Cobertura de alcantarillado (Censo)			■			■	
Penetración de banda ancha							
Déficit cuantitativo de vivienda (Censo)			■	■			
Déficit cualitativo de vivienda (Censo)			■	■			
Tasa de cobertura neta en educación preescolar	■			■	■		
Tasa de cobertura neta en educación primaria	■			■	■		
Tasa de cobertura neta en educación secundaria	■			■	■		
Tasa de cobertura neta en educación media	■			■	■		
Tasa de Mortalidad		■	■	■	■		■
Tasa de Mortalidad Materna		■	■				
Tasa de Fecundidad	■	■					
Tasa de Mortalidad Infantil		■	■	■			
Homicidios * 10000 hab	■			■			
Hurtos * 10000 hab	■			■			

Los pasos para determinar el impacto de las inversiones en los indicadores son los siguientes:

Una vez seleccionado un caso particular de evaluación de impacto, se procede a un acopio de toda la información pertinente que consiste en la conformación de un objeto tabular con la información asociada al *c-ésimo* ejercicio de evaluación. Por lo tanto, se construye una base de datos depurada que contiene información asociada al monto de la inversión, el indicador de desarrollo municipal y la lista de controles pertinentes para un corte transversal de todos los municipios de Colombia.

El objeto tabular conformado contiene la información mínima suficiente para la *c-ésimo* evaluación de impacto. No obstante, en su formato original la variable de inversión requiere un procesamiento específico. La teoría de evaluación de impacto está diseñada para problemas que admitan una representación a través de un tratamiento binario, es decir, donde se pueden identificar claramente los grupos de control y tratamiento. Teniendo en cuenta que el monto de inversión es un valor real continuo, se aplicó una discretización que permite identificar varios estratos de inversión. Por lo tanto, se considera que las entidades territoriales pueden ser expuestos a diferentes niveles de inversión, lo cual permite plantear relaciones dinámicas de control/tratamiento entre sucesivos niveles de inversión. Esta discretización satisface un criterio de optimalidad estadística, consistente en minimizar la varianza dentro de los grupos o clústeres conformados por cada nivel de tratamiento.

El emparejamiento busca que la distribución empírica de las características de los municipios tratados y no tratados sea lo suficientemente similar, como para justificar que las diferencia observadas en el indicador de desarrollo se atribuya únicamente a los diferentes niveles de inversión. En este caso se aplicó un algoritmo de emparejamiento sobre la toda muestra, procurando que las distribuciones de las características de los municipios sean similares en todos los niveles de tratamiento.

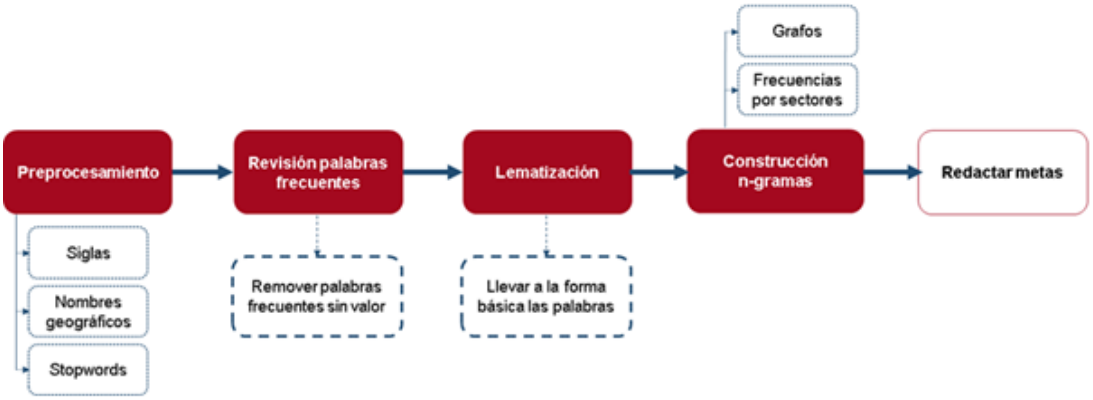
Frecuencias de N-gramas en las Metas de los Planes de Desarrollo Territorial

Los Planes de Desarrollo Territorial constituyen la principal herramienta de planificación y comunicación a partir de las cual las autoridades de cada entidad territorial establecen cuáles serán las acciones concretas por desarrollar dentro de su plan de gobierno, lo que los hace fundamentales para la transparencia y el control democrático y la preservación de la visión de gobierno.

La alta variabilidad en la formulación de las metas territoriales implica una alta dificultad para su seguimiento y evaluación por parte de otros organismos gubernamentales. Esta parte del proyecto busca el establecimiento de unas metas tipo por sector, que sirvan de orientación a la formulación de las metas territoriales, a través de plantillas básicas sirvan para estructurar su formulación. El insumo fundamental para construir estas plantillas son los rankings de *n-gramas* frecuentes por sector de inversión.

Los *n-gramas* son las unidades mínimas constituyentes de un texto que pueden ser objeto de análisis numérico. Consisten en un determinado número de palabras que conforman una misma entidad. Los n-gramas son las partículas fundamentales a partir de las cuáles las entidades territoriales interpretan y dan sentido a su experiencia histórica, y a las cuáles recurren para plantear soluciones dentro de un plan de gobierno coherente. Por ejemplo, el trigrama *reparación de víctimas* es especialmente representativo, tanto de la experiencia histórica que ha tenido un municipio, como de las acciones que consideran fundamental implementar en su política pública inmediata, para dar una síntesis a dicha experiencia histórica.

Desde un punto de vista metodológico, se asume una perspectiva de minería de texto. Es decir, asumimos que el conjunto de metas territoriales conforma una base de datos textual o *corpus de documentos* que se transforma en el insumo de análisis estadístico. En este caso, cada documento consiste en una meta territorial, la cual contiene la cadena de texto asociada a la descripción en lenguaje natural de la meta producida por la autoridad territorial competente.



En cualquier ejercicio de minería de textos, el primer paso es adoptar técnicas usuales de preprocesamiento del corpus, para que los textos adquieran una representación sistemática que facilite el análisis. Los documentos originales son afectados por una serie de normalizaciones (*convertion to lowercase*, *white-space-trimming*, *punctuation-removing*). También son descartadas las palabras frecuentes del idioma que no aportan ninguna intuición sobre el contenido semántico del documento (*stopword removal*). Además, también fue retirada una lista de palabras determinadas por un criterio humano (*human expert*), las cuáles eran irrelevantes el contexto de este trabajo (siglas, nombres geográficos). Para facilitar la sistematización del procesamiento estadístico, finalmente las palabras son contraídas a sus raíces fundamentales (*stemming*).

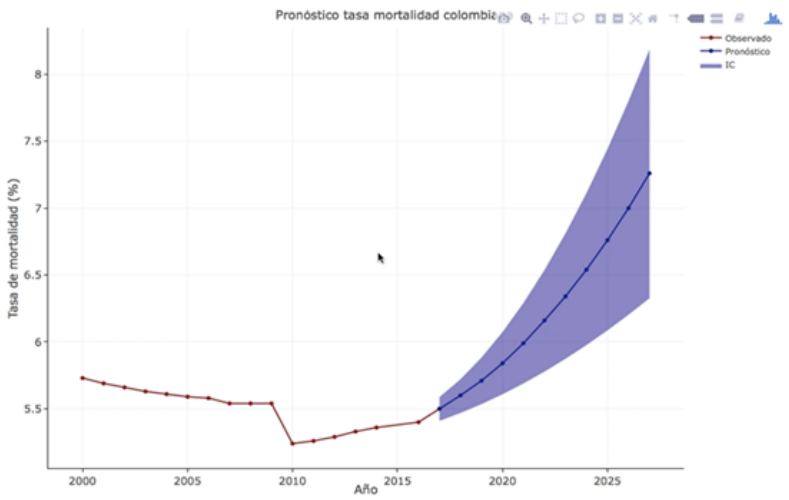
Siglas	Nombres geográficos	Stopwords	Palabras frecuentes
IDECA	colombia	de	implementar
FAMIS	república de colombia	la	programa
IDPAC	amazonas	que	proyecto
SIDAP	antioquia	el	plan
SIGOB	arauca	en	apoyar
COTSA	atlántico	y	población
ZODES	bogotá	a	público
IFTDH	bolívar	los	estrategia
PRAUS	boyacá	del	acción
RCLPD	caldas	se	proceso

Todas las anteriores rutinas contribuyen a obtener una representación más sistemática de una base de datos textual. A continuación, el corpus normalizado es particionado en una lista de bases de datos auxiliares que responden a la estructura sectorial de las metas territoriales. Finalmente se estiman los rankings de frecuencias de n-gramas por cada uno de los sectores.

Resultados

Pronóstico de Variables

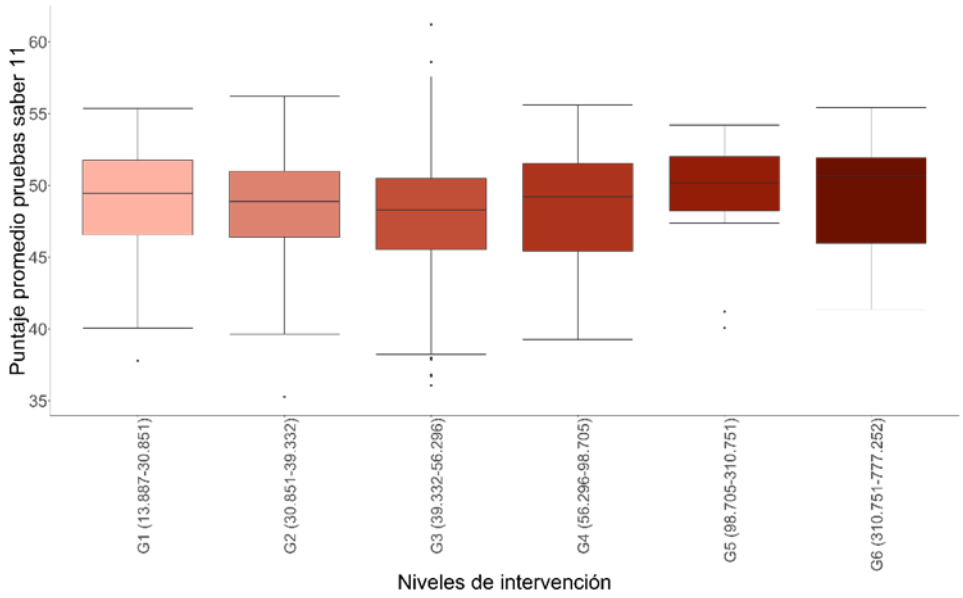
En el contexto de este proyecto se desarrolló un script para estimar y presentar el pronóstico de cualquier serie de tiempo considerada, a partir del modelo de Regresión Polinomial Local. Se presenta como ejemplo el resultado de aplicar la técnica de regresión polinomial local a la serie de tiempo de la Tasa de Mortalidad colombiana.



Los pronósticos parecen más verosímiles a corto que a mediano y largo plazo, debido a que este tipo de modelo trata de privilegiar la información local pero no pretende aproximar el comportamiento tendencial (de largo plazo) del proceso subyacente. Por ejemplo, la Tasa de Mortalidad Colombiana ha tenido un ligero resurgimiento en los últimos 10 años, lo cual logra ser aproximado por el modelo estimado y se refleja en el pronóstico de los años siguientes. Por otra parte, el pronóstico a 10 años no parece creíble, ya que implica un crecimiento sostenido de la Tasa de Mortalidad hasta un máximo histórico que no se ha observado en 25 años. La función de pronóstico asociada al modelo de Regresión Polinomial Local estaría extrapolando el comportamiento local sobre los pronósticos anteriores y no tiene recursos para aproximar el comportamiento de tendencia de muy largo plazo.

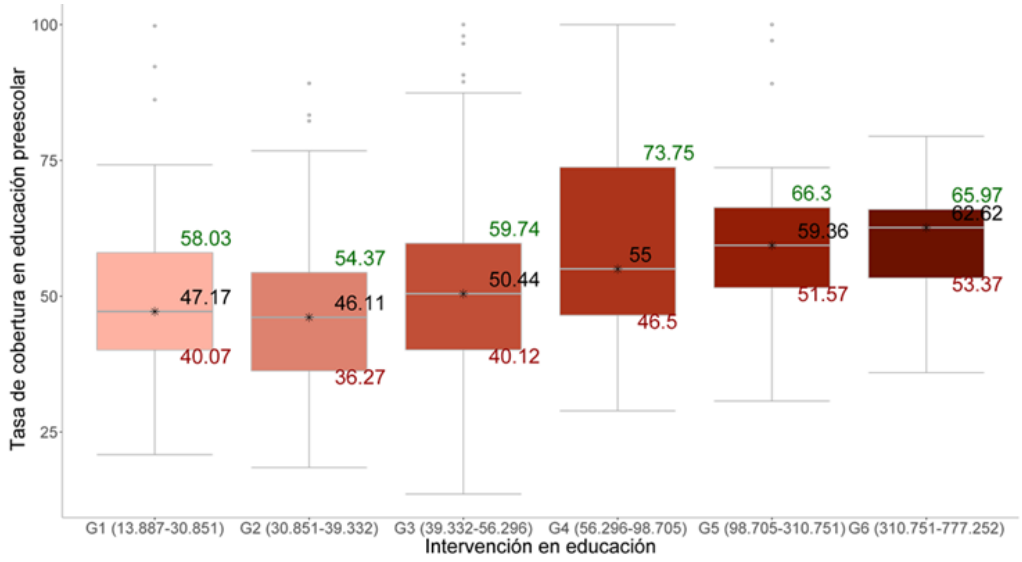
Estimación del Impacto de las Inversiones

La estimación a partir del modelo propuesto permite cuantificar el impacto de la inversión sobre el indicador de desarrollo municipal. Se recurrió a un simple modelo de medias que permite comparar el promedio muestral del resultado de interés entre niveles sucesivos de tratamiento.



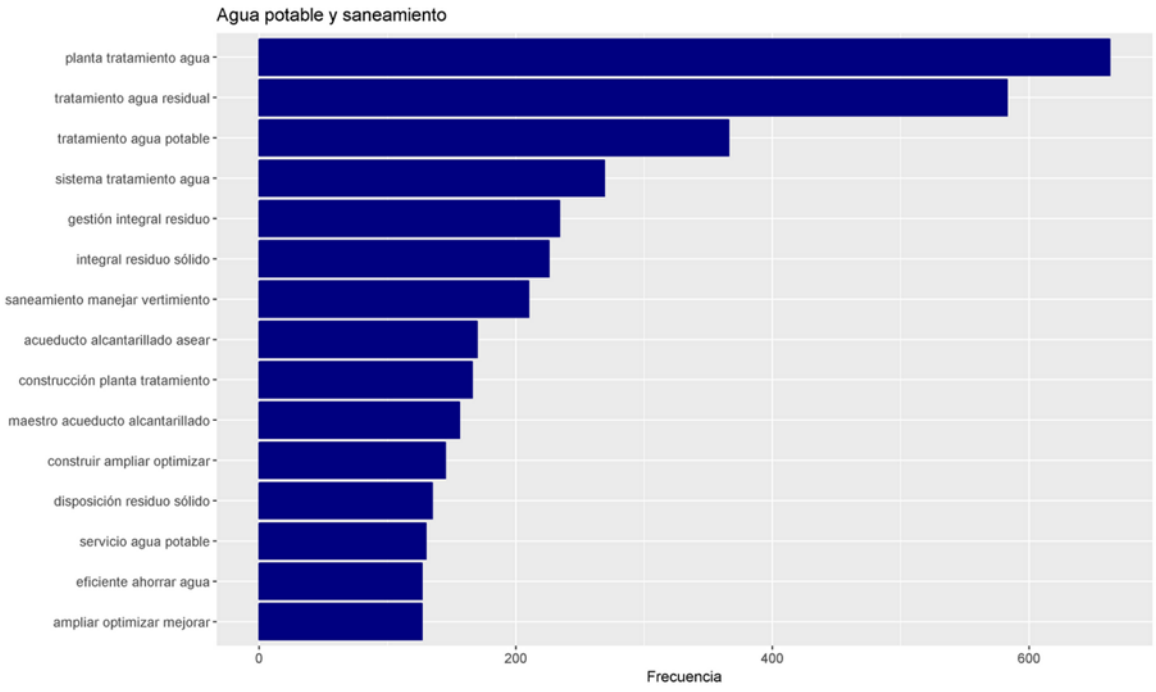
La metodología estadística propuesta permite estimar relaciones aproximadamente causales entre el monto de la inversión aplicada a las entidades territoriales y la evolución de indicador de desarrollo municipal. Esta metodología permite estimar la diferencia en el indicador para niveles sucesivos de tratamiento en una

muestra balanceada, donde las unidades de control y tratamiento son similares en sus características observadas por virtud del algoritmo de emparejamiento. La metodología desarrollada fue ilustrada para un caso particular de evaluación de impacto consistente en la inversión en educación.



Frecuencias de N-gramas en las Metas de los Planes de Desarrollo Territorial

El resultado de la operación estadística son los rankings de *n-gramas,* junto con su respectiva estructura sectorial. Se construyeron rankings de frecuencias de términos para unigramas, bigramas, trigramas, cuatrigramas y pentagramas, por cada uno de los 19 sectores que conforman los Planes de Desarrollo Territorial. La siguiente gráfica presenta un ejemplo de estos resultados, los trigramas más frecuentes para el sector de “Agua Potable y Saneamiento”.



Conclusiones

1. Se planteó una metodología para el pronóstico de las series de tiempo de las variables en la base Terridata. Esta metodología tiene mejor desempeño a horizontes de tiempo cercanos debido a su naturaleza de capturar el comportamiento local de la serie.
2. La metodología puede ser utilizada para elaborar pronósticos óptimos dado un conjunto de información univariada a cortos horizontes de pronóstico para cualquier variable considerada. Por lo tanto, es posible afirmar que este ejercicio aporta la disponibilidad de pronósticos fiables para las 350 variables de Terridata a un horizonte de pronóstico de 2 años.
3. Podría ser deseable indagar por metodologías que impliquen una estrategia explícita para representar el comportamiento tendencial de las series de tiempo, de forma que se pueden elaborar pronósticos óptimos de mediano y largo plazo para las variables presentes en Terridata.
4. La metodología de evaluación de impacto desarrollada se puede convertir en un importante insumo para la formulación de política pública basada en la promoción de la eficiencia administrativa en la asignación del gasto.
5. Es necesario aún analizar cuál es la estructura temporal relevante a cada ejercicio de evaluación de impacto, para incorporar este parámetro dentro de las consideraciones del proyecto.

Socialización

Los resultados de este proyecto fueron presentados internamente a la Dirección de Desarrollo Digital y la Dirección de Descentralización y Desarrollo Regional.