COLOMBIA POTENCIA DE LA VIDA

Departamento Nacional de Planeación - **DNP**

REG COL 3.0

Ampliación del periodo disponible (1991-2021), mejoramiento del modelo de clasificación sectorial, medición de la restrictividad e inclusión de métricas de complejidad de texto

IMPRENTA NACIONAL DE COLOMBIA

Observatorio de Mejora Normativa

REGCOL 3.0: ampliación del periodo disponible (1991-2021), mejoramiento del modelo de clasificación sectorial, medición de la restrictividad e inclusión de métricas de complejidad de texto.

Departamento Nacional de Planeación

Dirección General
Jorge Iván González

Subdirección General de Prospectiva y Desarrollo Nacional
Juan Miguel Gallego Acevedo

Dirección de Gobierno, Derechos Humanos y Paz
Lina María Valencia Ordóñez

Subdirección de Gobierno y Asuntos Internacionales
Maria Fernanda Fuentes Tuta

María Jimena Padilla Berrio
Lewis Enrique Polo Espinosa
Jacobo Campo Robledo
José Libardo Mejía Ciro
Valentina Porras Ocampo – diseño y diagramación

Unidad de Científicos de Datos
Hernan David Insuasti Ceballos

Apoyo editorial
Oficina Asesora de Comunicaciones
Jefe
Diana María Bohórquez Losada
Correctora de estilo
Carmen Elisa Villamizar Camargo

Imprenta Nacional de Colombia

Gerente General
Leonor Árias Barreto

Oficina Asesora de Planeación
Nassier Arenas Nuñez

Oficina de Seguridad Jurídica
Diego Insuasty Mora

Calle 26 # 13-19 Bogotá, D. C.
PBX: 60-1 381 5000

Contenido

01.

Introducción

El Observatorio de Mejora Normativa (OMN) de la Dirección de Gobierno, Derechos Humanos y Paz tiene como propósito garantizar el registro estadístico de la producción normativa nacional, dimensionar su magnitud y estudiar los posibles efectos económicos producidos por el nivel de regulación en Colombia. Con base en el principio regulador enunciado se desarrolló el proyecto REGCOL, cuyo propósito principal consiste en la consolidación de una base de datos que permite identificar el flujo de regulación al que se enfrentaron los sectores productivos en Colombia. La base de datos se desarrolla tomando el *Diario Oficial* de la Imprenta Nacional de Colombia como fuente de información oficial de normas en Colombia. Por otro lado, el desarrollo metodológico, teórico y tecnológico fue adelantado por parte de los grupos de Mejora Regulatoria y de Científicos de Datos del Departamento Nacional de Planeación.

En su primera versión se empleó como punto de partida el proyecto RegData del Mercatus Center de la Universidad de George Washington de Estados Unidos quienes analizan la regulación mediante técnicas de procesamiento de lenguaje natural, herramientas de minería de texto y aprendizaje de máquinas, se analizaron los textos regulatorios de los decretos y leyes publicadas en el *Diario Oficial* de Colombia. El proyecto REGCOL realizó la primera publicación en el año 2020 con los resultados encontrados por el modelo de clasificación del corpus regulatorio en textos sustanciales o no; los cuales posteriormente fueron clasificados en los sectores económicos a los que impacta. También se presentó el primer análisis de la tendencia del número de palabras y el número de expresiones vinculantes contenidas en los decretos y leyes para el periodo entre 1991 y 2014.

En este documento se presenta la actualización de la base REGCOL para su versión 3.0, y las mejoras de la versión actual respecto a las 1.0 y 2.0 se centran en cuatro aspectos fundamentales. En primer lugar, se actualizó la información hasta el año 2021; de tal forma, la base de datos comprende el flujo regulatorio de decretos y leyes entre 1991 y el 2021. Al igual que en la versión 2.0, para el periodo entre el 2014 y el 2021 se tiene disponible en formato de Word la totalidad del *Diario Oficial*

y no cada regulación en un archivo individual como era el caso en la versión 1.0. Desarrollando un modelo de separación de texto que permitió tomar los diarios completos y separarlos por regulaciones individuales, formato requerido para los pasos posteriores de clasificación de textos regulatorios.

Así mismo, para esta versión solo se tuvo en cuenta el modelo de clasificación *Support Vector Machine,* por ser el que tuvo mejor ajuste y desempeño en la versión 2.0 del proyecto; por tal razón, se incluye una sección dedicada solo a ese modelo y los resultados obtenidos para la nueva fase. En tercer lugar, se estimó la proporción de expresiones vinculantes relevantes con el propósito de mostrar una mejor aproximación al nivel de restrictividad, mediante la probabilidad de asociación de los actos administrativos de carácter sustancial a cada uno de los sectores económicos. En último lugar, se construyeron unas métricas de complejidad de texto que clasifica en distintos grados lo complicado que es para una persona comprender los textos de actos administrativos de acuerdo con el nivel de escolaridad alcanzado. Estas métricas son: cuentas condicionales, Dale-Chall y entropía de Shannon.

Con base en lo anteriormente señalado, en el presente documento se procede a explicar en detalle el proceso de actualización y de mejora de los algoritmos de clasificación y se anexan estadísticas descriptivas de la actualización de la base de datos comparable con los resultados ya presentados en las versiones 1.0 y 2.0 de REGCOL; posteriormente, se procede a explicar las nuevas variables disponibles de expresiones vinculantes relevantes y las medidas de complejidad y se ofrece un análisis detallado de los resultados.

Proceso de actualización del proyecto REGCOL

Esta sección detalla el proceso de actualización del proyecto REGCOL y se destacan las mejoras respecto a la versión 2.0[1]. Se especifica el proceso de extracción de los documentos con las regulaciones a partir de los diarios oficiales completos, el procedimiento de preprocesamiento establecido de limpieza y depuración de los textos, se explica el modelo de clasificación de textos seleccionado en la versión 2.0 el cual se emplea otra vez en la versión 3.0 del proyecto. Finalmente, se explican los cambios hechos en la versión más reciente del proyecto.

2. 1 Extracción individual de regulaciones de los diarios

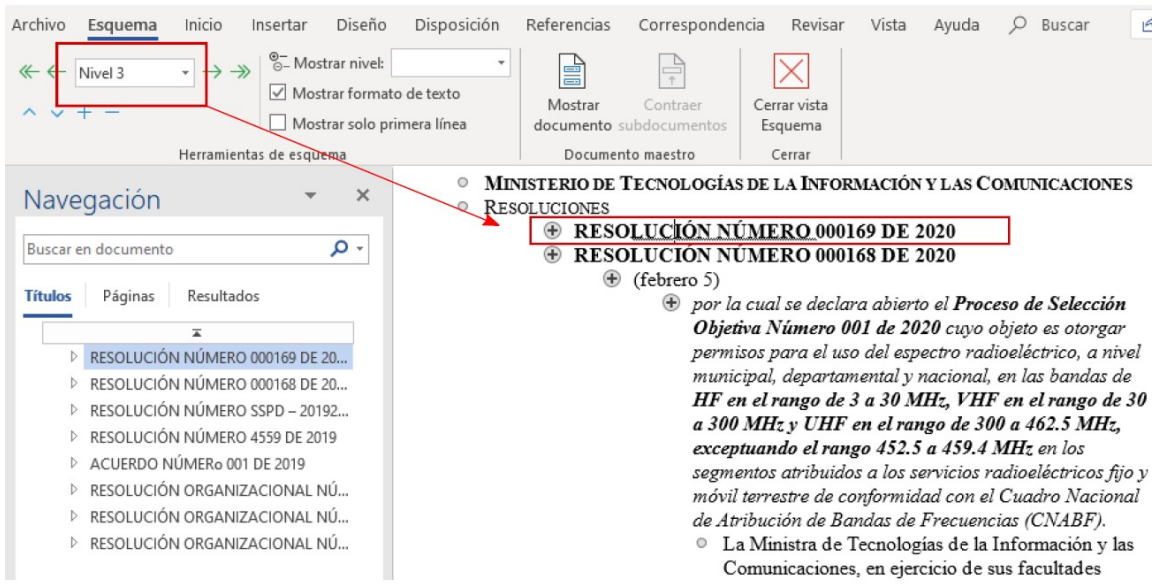
Al igual que el procedimiento descrito en la sección de metodología de construcción de REGCOL 2.0 (Observatorio de Mejora Normativa, 2021), en la presente versión del proyecto se tiene como fuente de información adicional los diarios completos actualizados al 2021; es decir, los decretos y leyes publicadas en el *Diario Oficial* desde el año 2014 hasta 2021 en formato Word[2] para la versión 3.0.

En la **figura 2-1** registra la estructura del formato Word entregado por el *Diario Oficial*, el cual presenta la estructura de título Nivel 3, el cual permite separar las diferentes regulaciones en el periodo bajo estudio. Con el objetivo de hacer el preprocesamiento de los textos se consolidó un código en el software Python que permite identificar la estructura de Nivel 3 y separa las diferentes regulaciones en una lista conformada por textos planos cuya distribución permite garantizar la estructura necesaria por los métodos de análisis de texto que permiten su clasificación en los diferentes sectores económicos.

1 REGCOL 2.0 - Ampliación del periodo disponible (1991-2019) y mejoramiento del modelo de clasificación sectorial: https://colaboracion.dnp.gov.co/CDT/ModernizacionEstado/MyE/REGCOL_2.0.pdf

2 Este formato lo denomina la Imprenta Nacional de Colombia como el “enlace” del *Diario Oficial.* Esta versión se utiliza para la organización previa a la diagramación y publicación final. Puede haber diferencias entre el “enlace” y la versión final publicable en formato PDF; sin embargo, estas son poco frecuentes y, por ende, no representan un problema para el proyecto.

Figura 2-1.
Visualización de organización por niveles del Diario Oficial en formato Word



Fuente: Unidad de Científicos de Datos (UCD) del Departamento Nacional de Planeación (DNP).

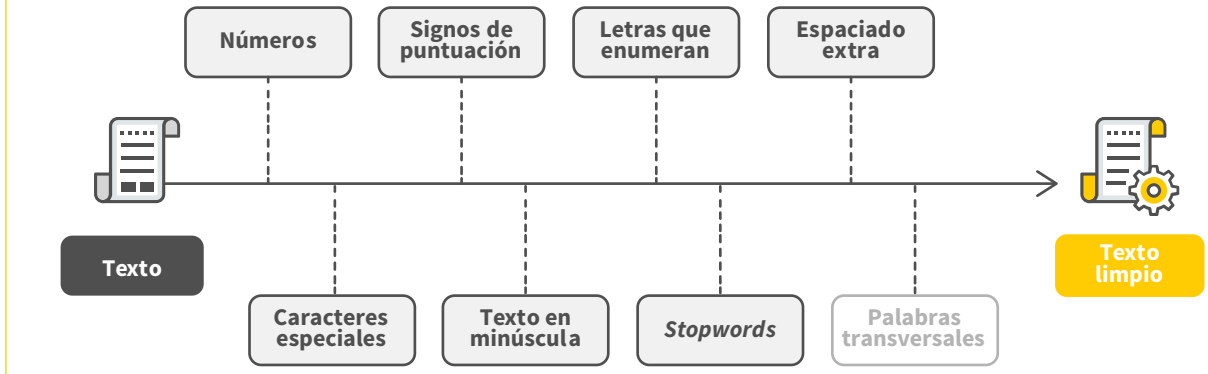
2. 2 Limpieza y extracción de características de los textos regulatorios

El preprocesamiento de los textos permite la extracción del ruido presente en los documentos de las regulaciones que no permite capturar la generalidad de su contenido con el fin de maximizar la capacidad del método de clasificación de textos (Grimmer *et al*., 2022; Hvitfeldt & Silge, 2021). En la actualidad existe un gran conjunto de metodologías de preparación de textos para el posterior uso de algoritmos de aprendizaje de maquina *(Machine Learning)* los cuales dependen del tipo de documento por analizar (Aggarwal, 2018; Evans & Aceves, 2016). Por tal motivo, se expone la estructura de preprocesamiento empleada sobre los textos de las regulaciones con el propósito de construir el conjunto de corpus estandarizado que será el insumo en el proceso de extracción de las características de los textos y el algoritmo de clasificación seleccionado para su procesamiento.

2.2.1 Limpieza de texto

Al igual que en las versiones previas del proyecto se depuró cada uno de los textos (regulaciones) bajo la secuencia ilustrada en la figura 2-2 empleando la librería de análisis de texto *ConTexto* desarrollada por la UCD para *Phyton*. En esta etapa del procesamiento se eliminan un conjunto de estructuras de los textos previo a su caracterización final; se removieron signos de puntuación, caracteres no alfanuméricos, números, acentos —tildes, eñes— y espacios innecesarios. A continuación, se eliminaron un conjunto de palabras vacías (*stopwords*) o no informativas como los son artículos, preposiciones, nombres geográficos, nombres y apellidos de personas hispanos y sudamericanos, y, finalmente, se eliminaron las letras —palabras con menos de dos caracteres— y las palabras con baja frecuencia —que aparecen menos de tres veces en el documento—. En la última parte del proceso de estandarización del texto se eliminaron las palabras y expresiones específicas indicadas por el experto temático y el corpus se pasó a minúscula.

Figura 2-2.
Proceso de limpieza de texto mediante la librería ConTexto desarrollada por la UCD



Fuente: DNP-UCD.

2.2.2 Extracción de características

Al igual que en la versión previa del proyecto (REGCOL 2.0), cada corpus estandarizado (normativa individual) se transforma en una representación numérica para poder ser procesado por el computador mediante el algoritmo TF-IDF *(Term Frequency - Inverse Document Frequency)*. Se emplea esa metodología en particular debido a que los cálculos obtenidos en REGCOL 2.0 muestran que dicho enfoque produce mejores resultados que otras propuestas similares para la representación de los textos y en las etapas posteriores de clasificación.

El TF-IDF es una medida estadística utilizada para evaluar la importancia de una palabra para un documento en un conjunto de documentos; es decir, la trascendencia de la palabra aumenta proporcionalmente al número de veces que aparece en el documento (parte TF), pero se compensa con la frecuencia de la palabra en los otros documentos que forman parte del conjunto de entrenamiento (parte IDF).

2. 3 Métodos de clasificación en el procesamiento de lenguaje natural

Uno de los procedimientos más empleados en el aprendizaje de máquina es la clasificación supervisada. Es un enfoque se tiene un amplio conjunto de metodologías o modelos estadísticos —se resaltan los árboles de clasificación, la regresión logística clásica y multinomial, análisis discriminante, entre otros—, y las propuestas específicas de la inteligencia artificial (IA) —entre las más conocidas están las redes neuronales clásicas o bayesianas, inducción de reglas—.

Los métodos de clasificación bajo el enfoque de interés hacen uso de una partición de los corpus estandarizados (normativa individual) en dos grupos que serán empleados como entrenamiento o prueba. El conjunto de normativas estandarizadas que es empleada en el proceso de entrenamiento permite la estimación de los parámetros requeridos por modelo de clasificación mientras que el subconjunto de documentos de prueba es empleado en el proceso de validación del comportamiento del modelo estimado. Se utiliza el *muestreo aleatorio simple* (MAS) como procedimiento estadístico de selección de corpus normativo, con el fin de garantizar que cada uno de ellos solo pertenezca a uno de los grupos previamente enunciados.

Como resultado de la implementación de un método de clasificación se tendrá una variable dicotómica que tomará el valor de 1 (Sí) cuando el corpus normativo es clasificado en uno de los sectores económicos (tabla 2-1). Estos métodos causan dos tipos de error, los falsos positivos y negativos; es decir, en el caso de una variable

binaria que toma valores 0 y 1, habrá ceros que se clasifiquen incorrectamente como unos y unos que se clasifiquen incorrectamente como ceros. Se espera que al incluir más documentos de entrenamiento en los métodos de clasificación, la tasa de error de clasificación tienda a cero.

Tabla 2 - 1.
Sectores productivos CIIU (Clasificación Industrial Internacional Uniforme)

1 Agricultura, caza, forestal y pesca	6 Comercio, hoteles y restaurantes
2 Minería y extracción	7 Transporte, almacenaje y comunicaciones
3 Manufacturas	8 Finanzas, negocios y bienes raíces
4 Electricidad, gas y suministros de agua	9 Servicios personales, Administración pública, salud, educación
5 Construcción	10 Otros (administrativos, no sustanciales)

Fuente: elaboración propia con datos del Observatorio de Mejora Normativa del DNP (OMN-DNP).

A continuación, se presenta el modelo empleado en la clasificación de los corpus en los diferentes sectores económicos.

2.3.1 Modelos de clasificación

Al igual que en la versión anterior de REGCOL, se emplearon las metodologías más utilizadas en la actualidad como el conjunto de clasificadores por emplearse para REGCOL 3.0 con el objetivo de maximizar la probabilidad de correcta clasificación de los corpus tanto en el conjunto de entrenamiento como en la prueba. De nuevo, se tienen dos etapas separadas de clasificación en dos tareas por separado; la primera, clasificar las normativas como sustanciales o no sustanciales; y la segunda, clasificar las normativas sustanciales en los nueve sectores productivos o en el sector transversal tomado como el décimo sector para efectos prácticos (**tabla 2-1**).

También se consideraron los siguientes métodos de clasificación:

1. Random Forest
2. Support Vector Machine
3. K Nearest Neighbors
4. Multinomial Naives Bayes
5. Logistic Regression
6. Gradient Boosting

En la elección del mejor método de clasificación se usó el criterio de validación cruzada con 5 grupos. También en REGCOL 3.0 el clasificador de mejor desempeño fue el Support Vector Machine (SVM) con parámetros. El método obtuvo una precisión promedio del 82 %, mientras que el enfoque tuvo un *recall* del 92 %, valor que indica la minimización del riesgo de dejar por fuera una normativa sustancial. También se empleó como límite de decisión óptimo *(threshold)* para garantizar este objetivo.

2.3.2 Modelo de clasificación sectorial

Al igual que en REGCOL 2.0, en la segunda etapa del proceso de clasificación se entrena el modelo de clasificación por sectores. Para su ejecución se utilizó el conjunto de corpus etiquetados manualmente[3] por el equipo técnico del Observatorio de Mejora Normativa; la distribución de la clasificación manual se especifica en la **tabla 2-2**.

Tabla 2-2.
Distribución por sector de los corpus de normativas para entrenamiento del SVM

Sector	Muestras
1	57
2	33
3	66
4	24
5	33
6	44
7	46
8	119
9	82
10	124
Total	628

Fuente: elaboración propia con datos DNP-UCD.

3 Todos los documentos empleados pasaron por el proceso de limpieza y la codificación TF-IDF para la clasificación de los documentos en sustancial o no sustancial.

En la segunda etapa de clasificación se repite el mismo procedimiento utilizado para el modelo sustancial o no sustancial; es decir, una metodología de validación cruzada de cinco grupos y búsqueda heurística de los parámetros de cada estimador en cada iteración; la diferencia radica en que el método de clasificación es de la forma "uno versus el resto" *(OneVsAll)*. El clasificador de mejor desempeño fue de nuevo el modelo SVM con parámetros siguientes:
C = 3, ClassWeight = balanced, Decision Function Shape ="ovr", γ = 1, Kernel = "rbf".

2.3.3 Métricas de restrictividad y complejidad de textos

Al igual que en la versión anterior de REGCOL, como resultado del modelo de clasificación se estimaron las probabilidades de asociación de cada uno de los documentos regulatorios en los diferentes sectores económicos. Dichas probabilidades permiten la construcción de las nuevas métricas de expresiones vinculantes relevantes y del total de palabras relevantes, las cuales pretenden recoger la restrictividad asociada a un documento regulatorio y que puede involucrar a más de un sector productivo al mismo tiempo.

Por último, se identificaron e implementaron un conjunto de métricas de complejidad y claridad de textos como aproximación a factores intrínsecos de las regulaciones que puedan afectar su entendimiento y, por ende, el cumplimiento. Las métricas empleadas fueron:

- Métrica de cuentas condicionales
- Métrica de Dale-Chall
- Métrica de entropía de Shannon

Para las *cuentas condicionales*, dado que corresponden —como su nombre lo indica— a un conteo de expresiones condicionales, al igual que para las *cuentas de restricciones o palabras*, la variable también se construye teniendo presente la ponderación de asociación de cada texto con los diferentes sectores usando las probabilidades calculadas por el modelo de clasificación sectorial. En el capítulo 4 del documento se describen las métricas enunciadas y el análisis de los resultados obtenidos.

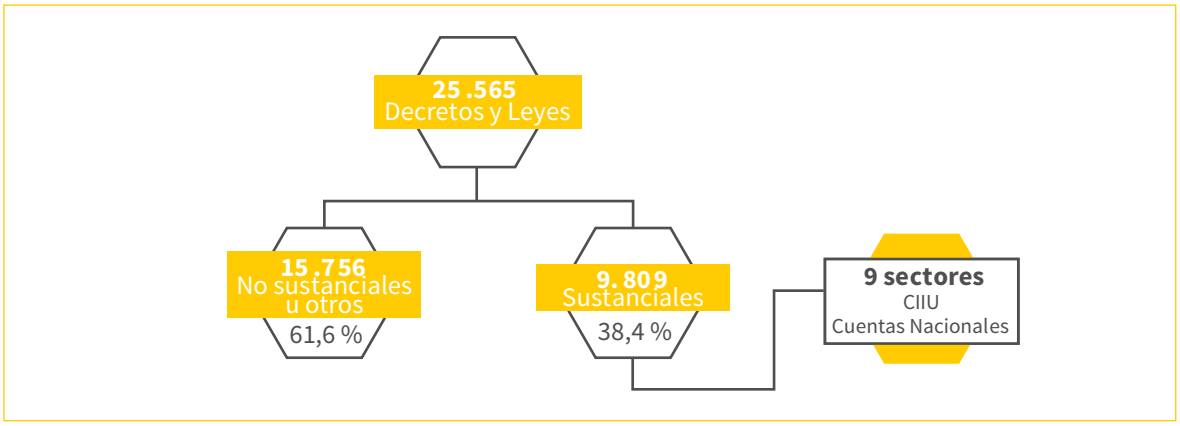
"...Como resultado del modelo de clasificación se estimaron las probabilidades de asociación de cada uno de los documentos regulatorios en los diferentes sectores económicos. Dichas probabilidades permiten la construcción de las nuevas métricas de expresiones vinculantes relevantes y del total de palabras relevantes, las cuales pretenden recoger la restrictividad asociada a un documento regulatorio y que puede involucrar a más de un sector productivo al mismo tiempo..."

Actualización del proyecto REGCOL para el periodo disponible (1991-2021)

Aquí se reúnen los resultados actualizados del proyecto para el periodo 1991-2021. En el REGCOL versión 3.0 se emplearon las mismas métricas consideradas en el documento técnico REGCOL 2.0: ampliación del periodo disponible (1991-2019) y mejoramiento del modelo de clasificación sectorial. La actualización tiene como objetivo presentar los cálculos ajustados de las métricas de interés para garantizar la comparabilidad de lo obtenido con los resultados ya encontrados en REGOCOL 2.0, condicionado al aumento del corpus regulatorio en REGCOL 3.0 dado por la ampliación del periodo temporal bajo análisis.

Con base en la descripción de la sección 2 del presente documento, respecto al proceso de actualización del proyecto REGCOL desde su segunda fase, se incluyen los principales resultados obtenidos durante proceso. Para comenzar el análisis se parte de que para el periodo 1991-2021 se recopilaron 25.565 textos reglamentarios de decretos y leyes expedidos en Colombia, de los cuales el 61,6 % fueron clasificados como *no sustanciales* u *otros* y el restante 38,4 % como *sustanciales* (**figura 3-1**). De esa muestra, se clasificaron sectorialmente solo las 9.809 regulaciones dentro de la categoría sustancial (**figura 3-1**). Por ello es evidente un crecimiento del 24 % en la producción de actos administrativos de carácter sustancial comparado con la fase 2.0 del proyecto.

Figura 3-1.
Descripción de la clasificación sustancial de los decretos y leyes publicados entre 1991 y 2021 en Colombia (nueve sectores)



Fuente: elaboración propia con datos OMN-DNP.

En relación con la agregación de la información por sector y año —consolidando las variables *número total de palabras y expresiones vinculantes*— arrojó un total de 313 observaciones correspondientes a treinta y un años y nueve sectores para la base consolidada de REGCOL 3.0, versión en también la que estarán disponibles las regulaciones *sustanciales* categorizadas en el sector *Otros*, por ser de carácter transversal en sectores diferentes, tal y como se analizó en la versión 2.0.

En consecuencia, se tiene que las regulaciones caracterizadas en nueve *sectores productivos* y la categoría *Otros* contienen en promedio 101.163 palabras y 270 expresiones vinculantes, lo cual implica que, en promedio, el 0,27 % de las palabras contenidas dentro de la regulación pueden llegar a implicar cumplimiento obligatorio **(tabla 3-1)**. Es decir, en los últimos dos años hubo un incremento de 0,04 puntos porcentuales (p. p.) en las palabras dentro de la regulación que pueden implicar cumplimiento obligatorio.

Tabla 3-1
Estadísticas descriptivas REGCOL 3.0 (nueve sectores y categoría Otros)

Variables	Media	Desviación estándar
Total palabras	101.163	139.917,7
Expresiones vinculantes	270	386,41
Vinculantes/Tot. palabras	0,0027	0,001
Años: 31 Sectores: 10 Observaciones: 313		

Fuente: elaboración propia con datos DNP-OMN.

Por otro lado, si se consideran las regulaciones direccionadas exclusivamente a los nueve sectores productivos, se encuentra que estos actos administrativos contienen, en promedio, 64.009 palabras y 172 expresiones vinculantes; ello implica que, en promedio, el 0,27 % de las palabras contenidas dentro de la regulación pueden implicar cumplimiento obligatorio dentro de estos sectores productivos **(tabla 3-2)**. Si bien la razón entre palabras vinculantes y total de palabras se demuestra una razón similar de palabras vinculantes y total de palabras, se puede apreciar que la categoría *Otros* incrementa el promedio observado tanto de total de palabras como de expresiones vinculantes. Lo anterior indica que los textos regulatorios dentro de esa categoría son más extensos que los clasificados sectorialmente, quizá por su carácter transversal.

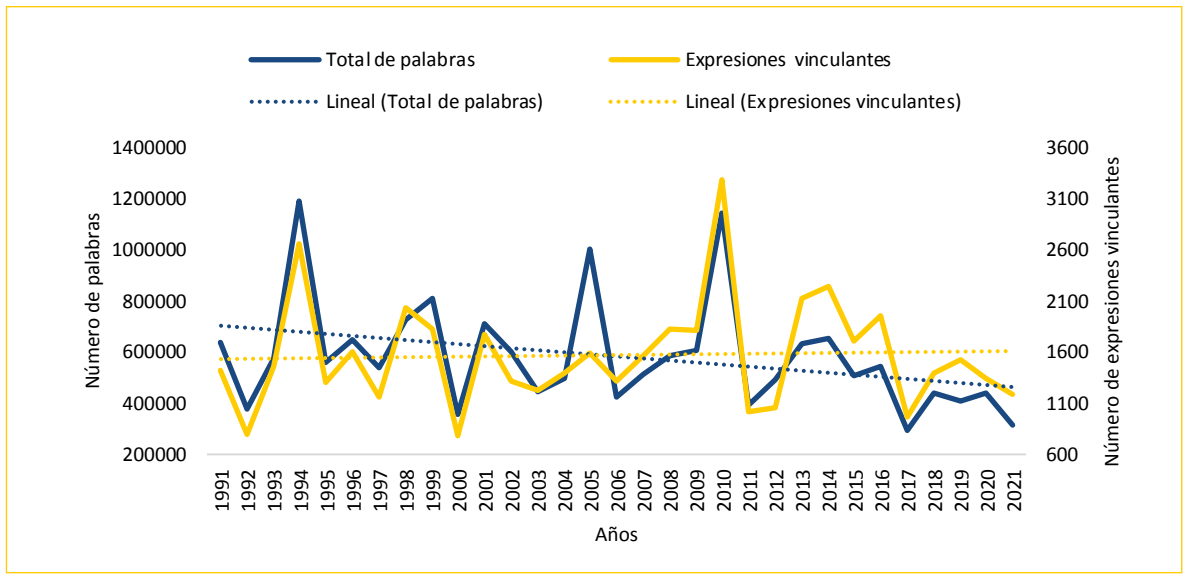
Tabla 3-2
Estadísticas descriptivas REGCOL 3.0 (nueve sectores)

Variables	Media	Desviación estándar
Total palabras	64.009	70.689,71
Expresiones vinculantes	172	198,64
Vinculantes/Tot. palabras	0,0027	0,001
Años: 31 Sectores: 9 Observaciones: 281		

Fuente: elaboración propia con datos DNP-OMN.

Iniciando con la evolución en el agregado del total de palabras y expresiones vinculantes de los actos administrativos, tal y como se ilustra en la **figura 3-2**, se observa que en todo el periodo de análisis tanto el total de palabras como de expresiones vinculantes responden a una misma tendencia (excepto en 2005); es decir, existe una consistencia y correlación entre dichas variables. Por su parte, al analizar los picos durante todo el periodo, se destaca que estos crecimientos, por lo regular, se han dado justo en el primer año de mandato de los presidentes, lo que cobra sentido teniendo en cuenta que se deben emitir actos administrativos con mayor cumplimiento u obligatoriedad. Resulta importante destacar que a medida que pasa el tiempo el número promedio de palabras en las regulaciones disminuye (**línea azul punteada**), mientras que el número de palabras vinculantes (cumplimiento obligatorio) aumenta (**línea amarilla punteada**).

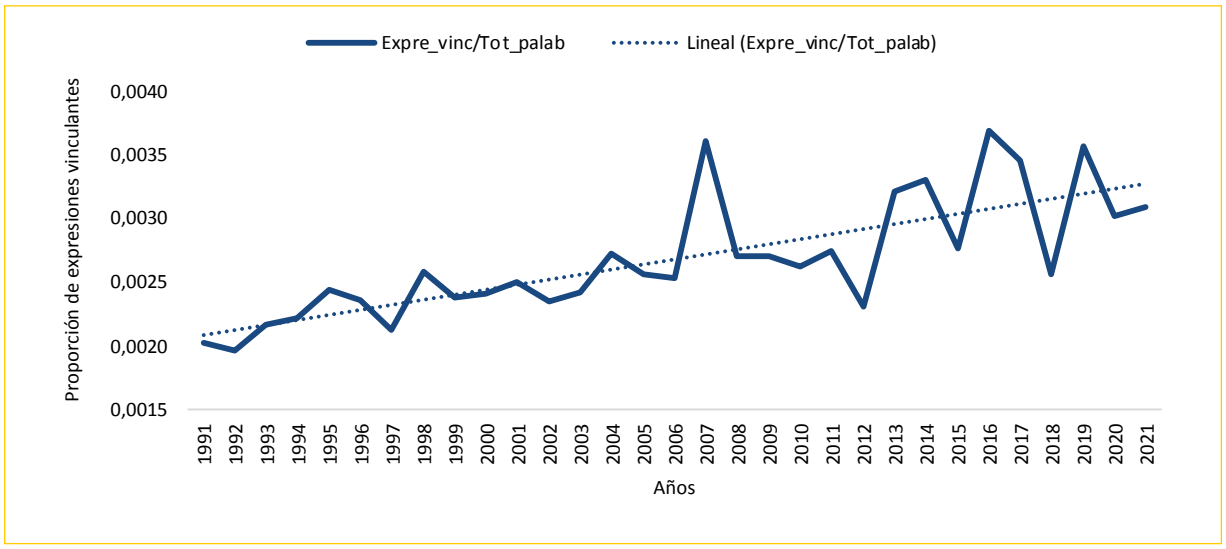
Figura 3-2.
Evolución agregada del total de palabras y expresiones vinculantes de decretos y leyes que regulan nueve sectores productivos en Colombia, 1991-2021



Fuente: elaboración propia con datos OMN-DNP.

Por otro lado, al analizar los patrones estructurales de la proporción de expresiones vinculantes respecto al total de palabras contenidas en los actos administrativos, se observa una tendencia creciente en la serie de 1991 hasta 2007 y solo hasta 2013 retoma su tendencia creciente en la evolución de la proporción. Por otro lado, al examinar la proporción de expresiones vinculantes respecto al total de palabras contenidas en la regulación se encuentra que, hasta 2017, iba en crecimiento la restrictividad u obligatoriedad de cumplimiento implícita dentro de las regulaciones **(figura 3-3)**.

Figura 3-3.
Evolución de la proporción de expresiones vinculantes respecto al total de palabras de decretos y leyes que regulan nueve sectores productivos en Colombia, 1991-2021.



Fuente: elaboración propia con datos OMN-DNP.

Ya presentado el comportamiento agregado del flujo de expresiones vinculantes y del total de palabras contenidas en los decretos y leyes clasificadas sectorialmente, se procede ahora a describir las particularidades de la regulación por sectores para el periodo 1991-2021.

Los sectores que, en promedio, se enfrentaron anualmente a un mayor flujo de palabras contenidas en la regulación específica fueron el de *comercio, restaurantes y hoteles* (142.785), el de *servicios financieros y empresariales* (130.310) y, en tercer lugar, el de *servicios comunales, sociales y personales* (81.764). De forma similar, los sectores que se enfrentaron, en promedio, a una mayor cantidad de expresiones vinculantes fueron el de *servicios financieros y empresariales* (439), el de *comercio, restaurantes y hoteles* (313) y, de tercero, *servicios comunales, sociales y personales* (215). Así mismo, es importante resaltar que la categoría *Otros*, como ya se mencionó, está compuesta por regulaciones que, en promedio, tienen mayor cantidad de palabras (432.067) y expresiones vinculantes (1.141) con respecto a los demás sectores productivos **(tabla 3-3)**.

En la **tabla 3-3** puede verse la gran heterogeneidad entre los valores tanto del flujo del total de palabras como del de expresiones vinculantes; esa característica implica que el flujo de regulación en los diferentes sectores no es una variable con una tendencia constante. Dicha variabilidad puede estar asociada a las dinámicas particulares que a lo largo del periodo 1991-2021 han requerido intervención regulatoria. Distinto de las versiones previas del proyecto, el sector con mayor variabilidad en los valores observados de regulación es el de *comercio, restaurantes y hoteles*, dado que las desviaciones estándar tanto del total de palabras como de las expresiones vinculantes son más numerosas en comparación con los demás sectores económicos.

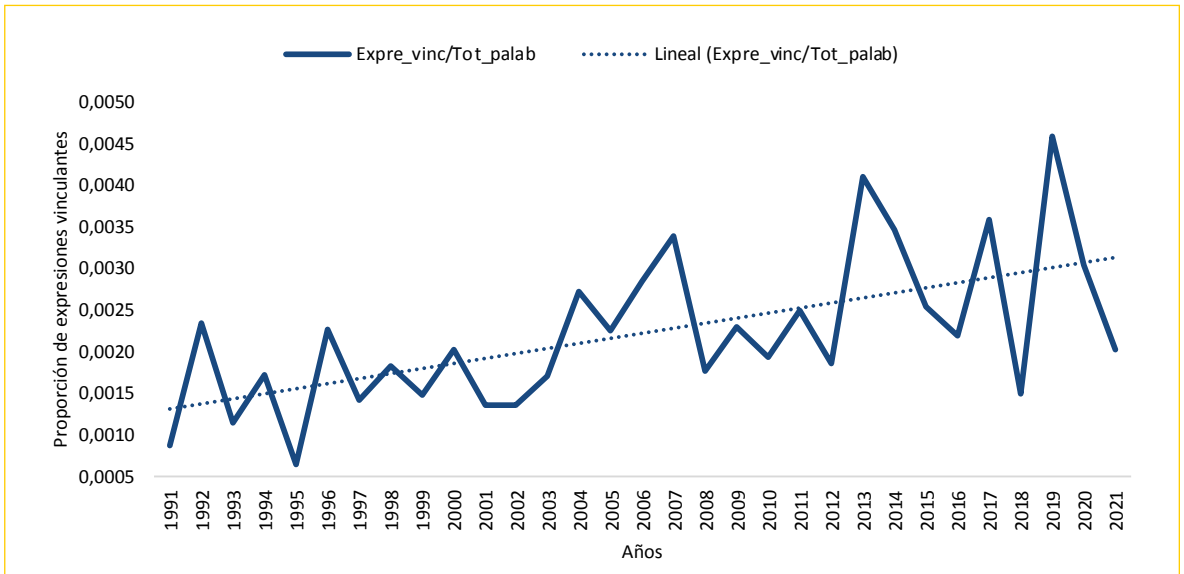
Tabla 3-3.
Estadísticas descriptivas REGCOL 2.0 (nueve sectores y categoría Otros)

Sector	Variable	Media	Desviación estándar	Mínimo	Máximo
Agropecuario, silvícola y pesquero	Total palabras	26.548,84	17.951,59	4.560	69.013
	Expresiones vinculantes	52,41	32,93	10	126
Minería	Total palabras	27.731,65	16.572,89	3.490	64.497
	Expresiones vinculantes	70,94	47,36	3	174
Manufacturas	Total palabras	42.704,69	25.026,22	6.856	101.951
	Expresiones vinculantes	110,38	81,17	20	393
Electricidad, gas y agua	Total palabras	12.220,26	9.512,98	310	37.818
	Expresiones vinculantes	35,61	35,17	0	146
Construcción	Total palabras	60.953,58	64.054,66	15.305	379187
	Expresiones vinculantes	192,58	193,30	54	1.097
Comercio, restaurantes y hoteles	Total palabras	142.784,91	110.323,56	8.465	485.394
	Expresiones vinculantes	313,72	217,14	16	768
Transporte y comunicaciones	Total palabras	48.211,97	53.496,25	716	246.831
	Expresiones vinculantes	114,19	111,86	6	519
Servicios financieros y empresariales	Total palabras	130.310,13	82.903,22	33.420	448.264
	Expresiones vinculantes	439,59	311,42	116	1.802
Servicios comunales, sociales y personales	Total palabras	81.764,38	33906,28	24.630	168.060
	Expresiones vinculantes	215,50	98,29	64	485
Otros	Total palabras	432.066,88	167.507,33	130.504	968.304
	Expresiones vinculantes	1.140,91	538,54	364	3090

Fuente: elaboración propia con datos DNP-OMN.

A continuación, se expone la proporción de expresiones vinculantes respecto al total de palabras contenidas en los decretos y leyes clasificados para cada sector. En primer lugar, la proporción de expresiones vinculantes respecto al total de palabras de las regulaciones del sector *agropecuario, silvícola y pesquero* presentó una tendencia creciente con dos picos pronunciados en los años 2013 y 2019; ello implica que mientras pasa el tiempo hay un mayor porcentaje de palabras vinculantes en este sector **(figura 3-4)**. En promedio, entre 1991 y 2021, el contenido restrictivo de la regulación del sector fue de un 0,24 %.

Figura 3-4.
Proporción de expresiones vinculantes sobre el total de palabras de las regulaciones, sector agropecuario, silvícola y pesquero

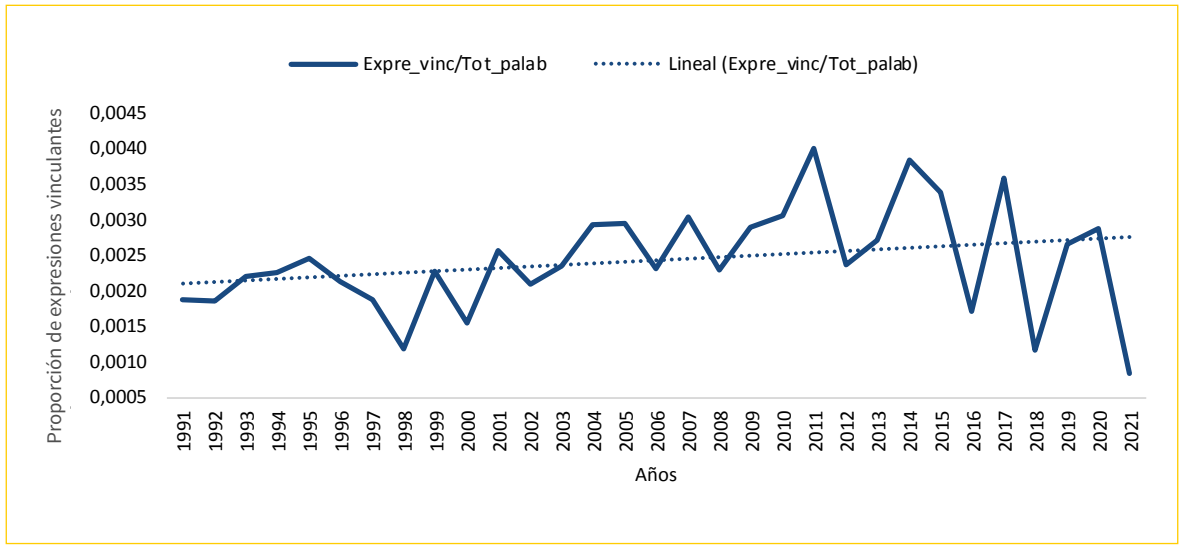


Fuente: elaboración propia con datos OMN-DNP.

La proporción de expresiones vinculantes respecto al total de palabras de las regulaciones del sector de *minería* presentó una tendencia creciente hasta el año 2011 y después decreció hasta el año 2021.

El mínimo observado se registró en el año 2021 cuando se ubicó en el 0,08 % de restrictividad **(figura 3-5)**. En promedio, entre 1991 y 2021, el contenido restrictivo de la regulación del sector fue de un 0,25 %.

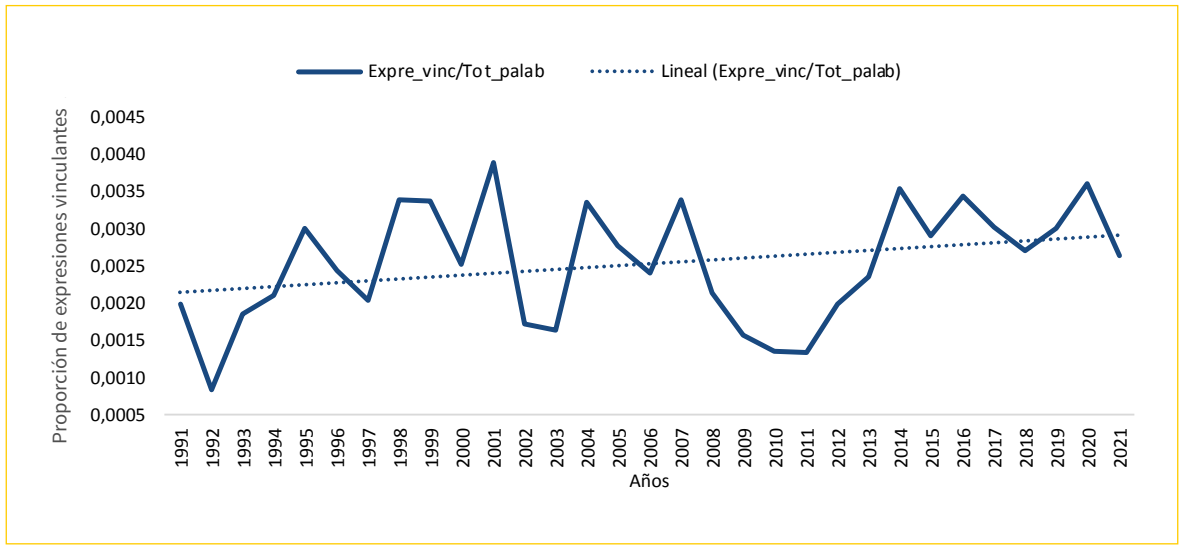
Figura 3-5.
Proporción de expresiones vinculantes sobre el total de palabras de las regulaciones, sector de minería



Fuente: elaboración propia con datos DNP-OMN.

La proporción de expresiones vinculantes respecto al total de palabras de las regulaciones del sector de *manufacturas* presentó una tendencia creciente durante el periodo bajo estudio. En lapso de 1991 a 2021 se observan tres grandes disminuciones en 1992, 2002 y 2010. El contenido restrictivo de la regulación del sector fue mínimo en el año 2010 cuando se ubicó en el 0,16 % de restrictividad (**figura 3-6**). En promedio, entre 1991 y 2021, el contenido restrictivo de la regulación del sector fue de un 0,29 %.

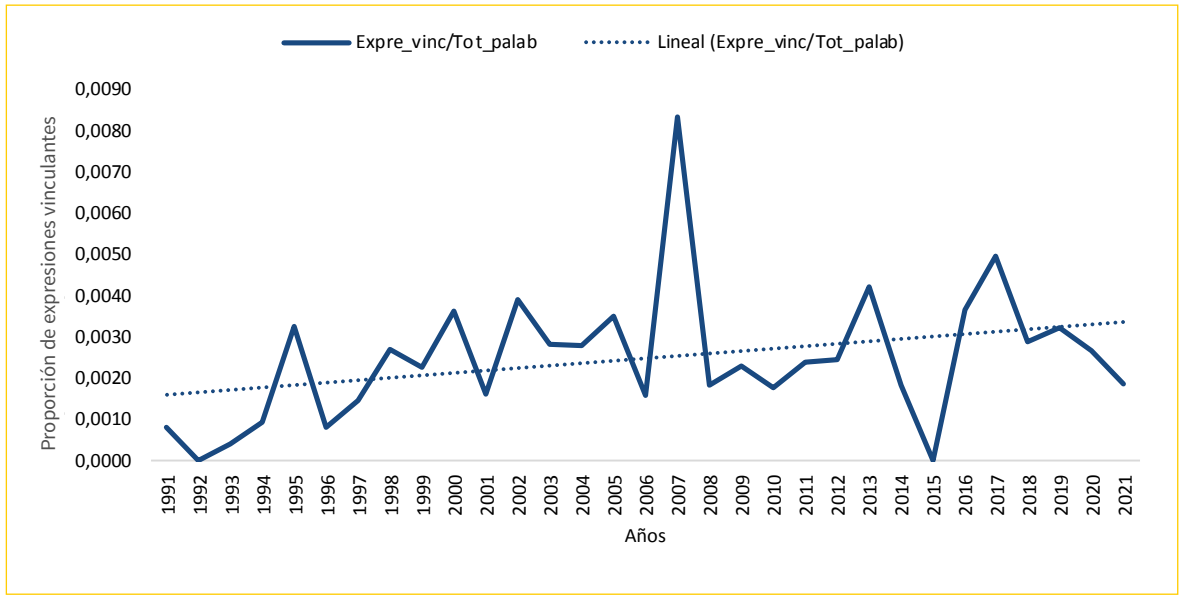
Figura 3-6.
Proporción de expresiones vinculantes sobre el total de palabras de las regulaciones, sector manufacturas



Fuente: elaboración propia con datos DNP-OMN.

Continuando con el análisis de los sectores económicos se determinó que la proporción de expresiones vinculantes respecto al total de palabras de las regulaciones del sector de *electricidad, gas y agua* tiene un patrón de tendencia creciente a lo largo de todo el periodo observado. Se observa que durante el periodo 1999 y 2005 la restrictividad de la regulación se mantuvo aproximadamente constante, pero se incrementó con notoriedad hasta alcanzar el máximo en el año 2007 **(figura 3-7)**. En promedio, entre 1991 y 2021, el contenido restrictivo de la regulación del sector fue del 0,27 %.

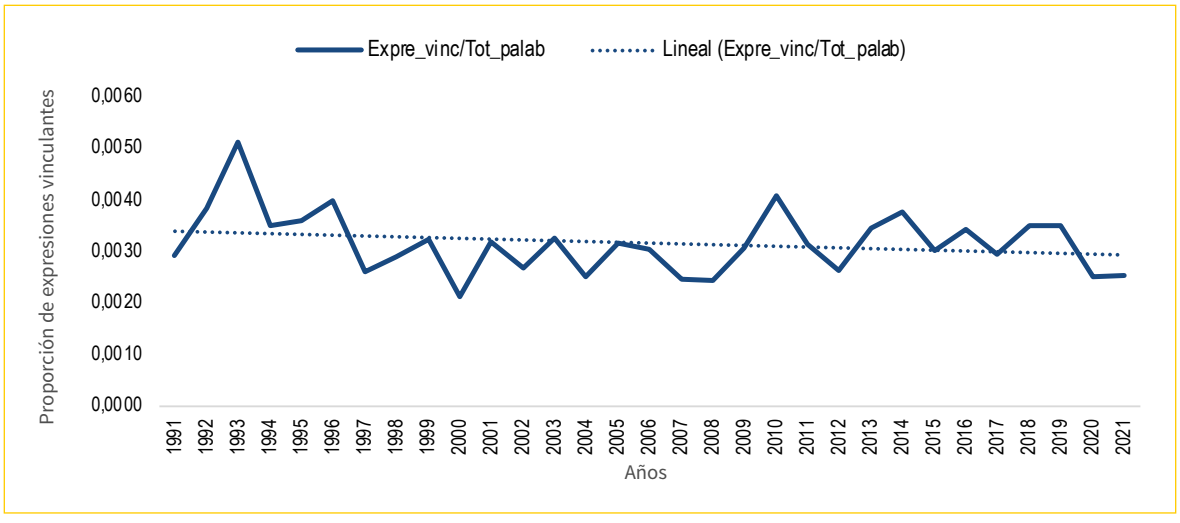
Figura 3-7.
Proporción de expresiones vinculantes sobre el total de palabras de las regulaciones, sector de electricidad, gas y agua



Fuente: elaboración propia con datos DNP-OMN.

Ahora, en el sector de *construcción*, la proporción de expresiones vinculantes respecto al total de palabras de las regulaciones del presentó una tendencia decreciente entre 1991 y 2021 **(figura 3-8)**. La tendencia contrasta con la alta variabilidad de las variables de regulación descritas antes; por ende, se observa que aunque la regulación varíe mucho de año a año, la vinculatoriedad parece disminuir con el tiempo. En promedio, entre 1991 y 2021, el contenido restrictivo de la regulación del sector fue del 0,33 %.

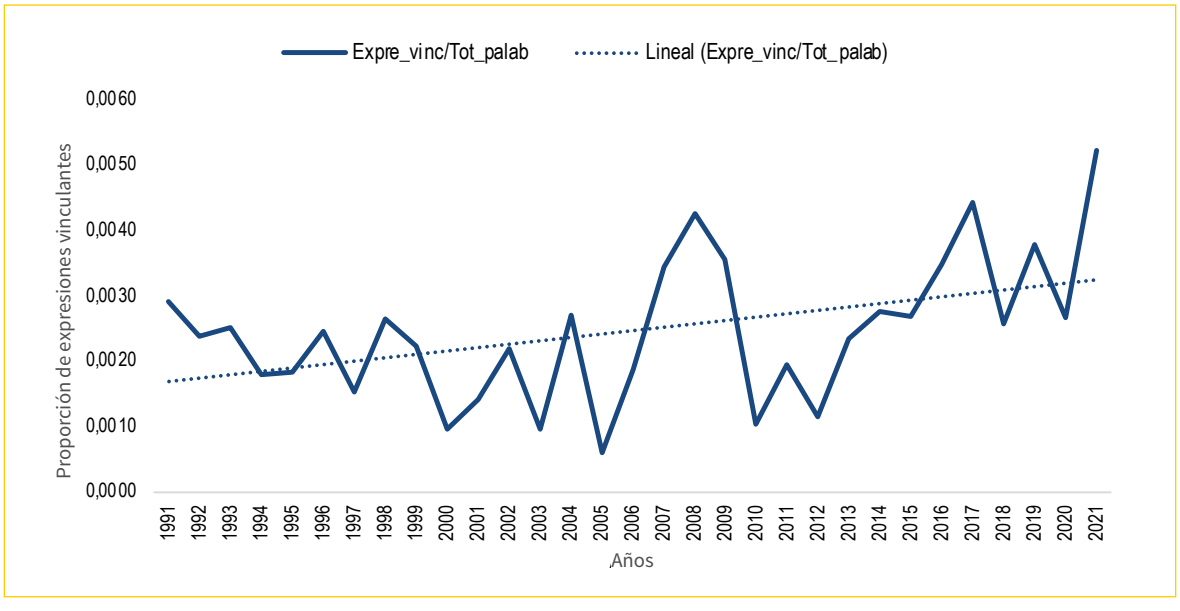
Figura 3-8.
Proporción de expresiones vinculantes sobre el total de palabras de las regulaciones, sector de construcción



Fuente: elaboración propia con datos DNP-OMN.

Al continuar el análisis de los sectores económicos se determinó que la proporción de expresiones vinculantes respecto al total de palabras de las regulaciones del sector de *comercio, restaurantes y hoteles* presenta una tendencia creciente en el periodo 2000 y 2021 **(figura 3-9)**; antes de ese periodo la tendencia era contraria (decreciente). En el periodo de crecimiento de la restrictividad se observa un gran pico en el año 2008 cuando llegó al 0,47 %. En promedio, entre 1991 y 2021, el contenido restrictivo de la regulación del sector fue de un 0,26 %.

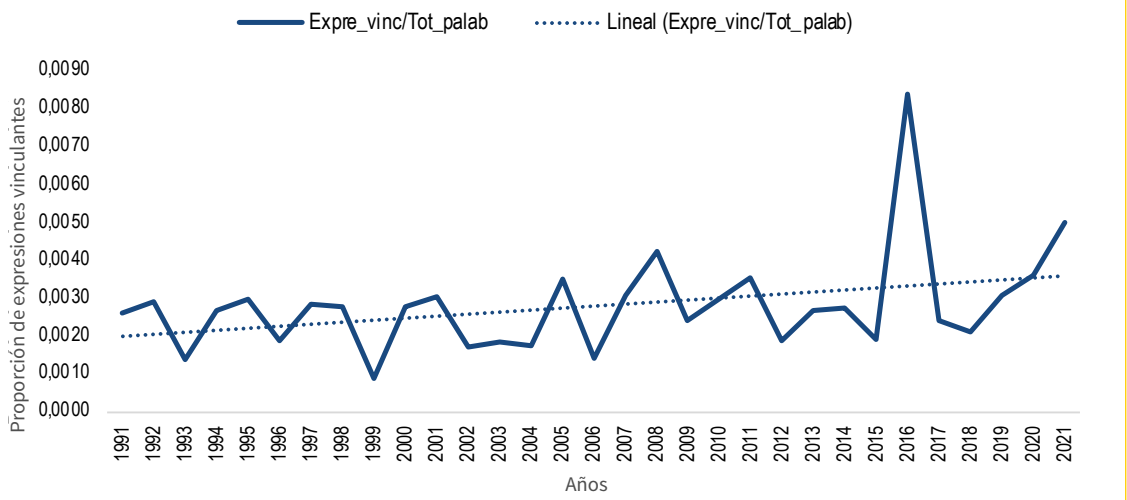
Figura 3-9.
Proporción de expresiones vinculantes sobre el total de palabras de las regulaciones, sector de comercio, restaurantes y hoteles



Fuente: elaboración propia con datos DNP-OMN.

Respecto al sector *transporte y comunicaciones* se puede observar que la proporción de expresiones vinculantes respecto al total de palabras de las regulaciones presenta una tendencia creciente y moderada durante el periodo analizado; también se puede apreciar un gran pico en el año 2016 cuando llegó al 0,84 % **(figura 3-10)**. La menor restrictividad se presentó en el año 1999 cuando se ubicó en el 0,09 %. En promedio, entre 1991 y 2021, el contenido restrictivo de la regulación del sector fue de un 0,30 %.

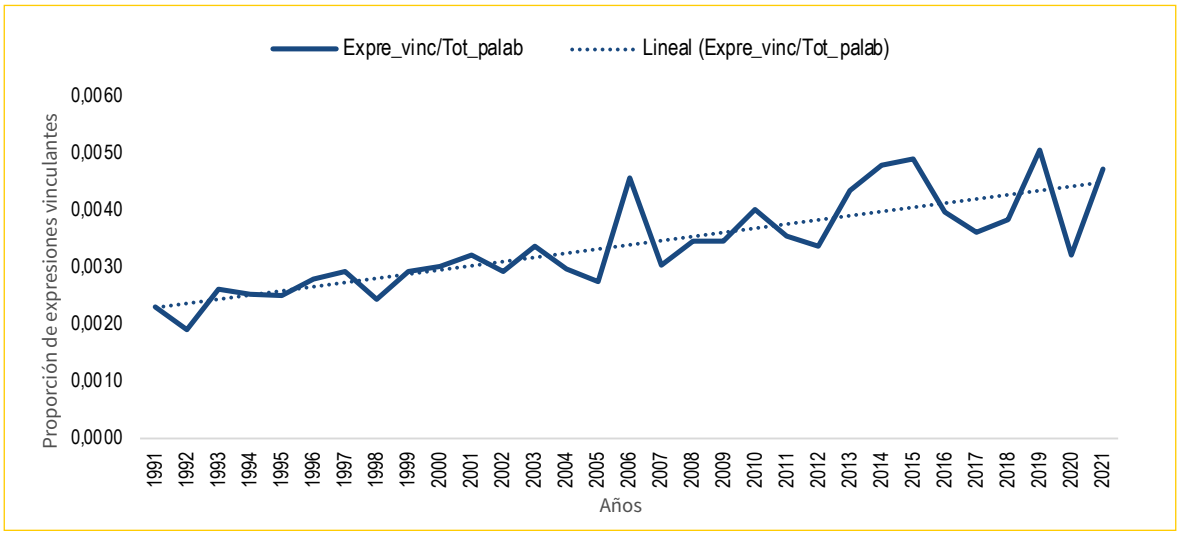
Figura 3-10.
Proporción de expresiones vinculantes sobre el total de palabras de las regulaciones, sector de transporte y comunicaciones



Fuente: elaboración propia con datos DNP-OMN.

En la **figura 3-11**, plantea la proporción de expresiones vinculantes respecto al total de palabras de las regulaciones del sector de *servicios financieros y empresariales*. Puede observarse una tendencia creciente durante todo el periodo en análisis, con un pico notorio en 2006 donde la restrictividad alcanzó el 0,48 %. En promedio, entre 1991 y 2021, el contenido restrictivo de la regulación del sector fue de un 0,36 %.

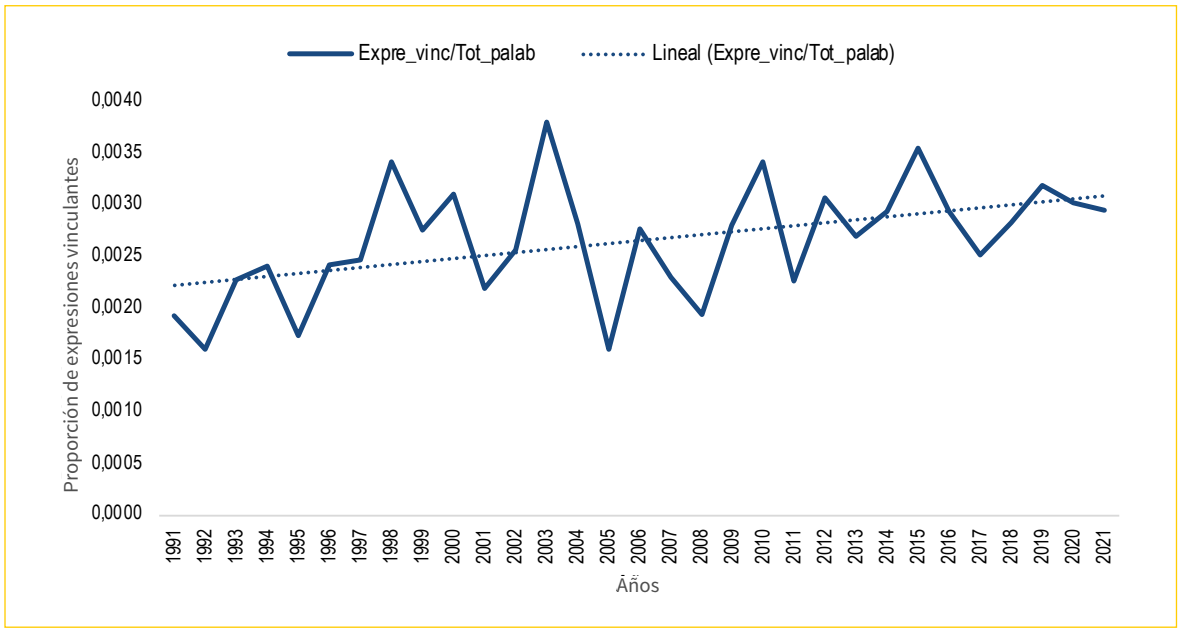
Figura 3-11.
Proporción de expresiones vinculantes sobre el total de palabras de las regulaciones, sector de servicios financieros y empresariales



Fuente: elaboración propia con datos DNP-OMN.

La proporción de expresiones vinculantes respecto al total de palabras de las regulaciones del sector de *servicios comunales, sociales y personales* presentó una tendencia creciente durante el periodo bajo análisis. También se puede observar un pico en la restrictividad en el año 2003 cuando llegó al máximo observado del 0,41 %, seguido de un mínimo de restrictividad en el 2005 que llegó al valor de 0,17 % **(figura 3-12)**. En promedio, entre 1991 y 2021, el contenido restrictivo de la regulación del sector fue de un 0,28 %.

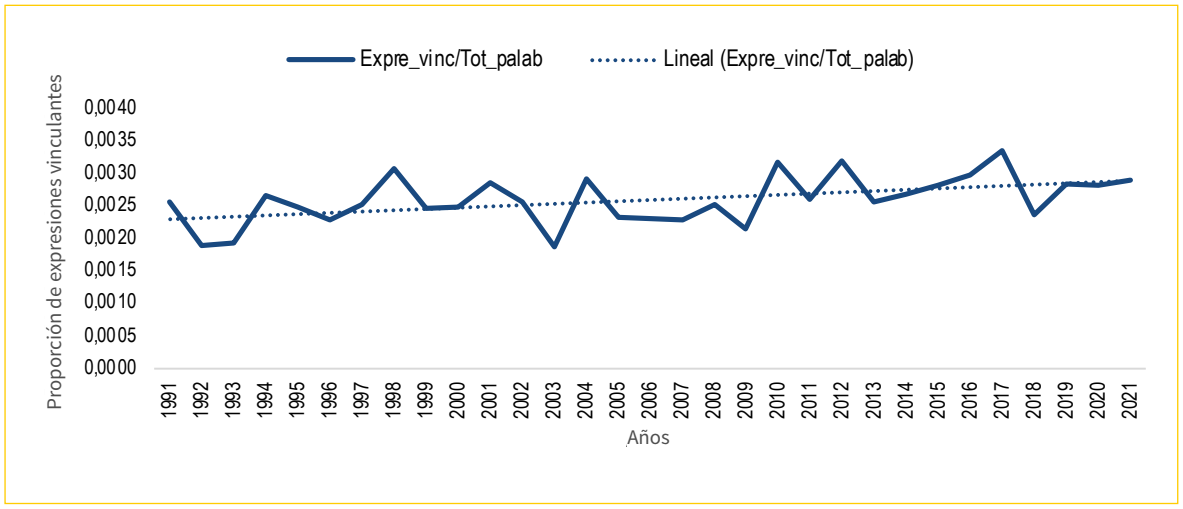
Figura 3-12.
Proporción de expresiones vinculantes sobre el total de palabras de las regulaciones, sector servicios comunales, sociales y personales



Fuente: elaboración propia con datos DNP-OMN.

La proporción de expresiones vinculantes respecto al total de palabras de las regulaciones de la categoría *Otros* presentó una tendencia casi que constante durante el periodo bajo análisis (**figura 3-13**). En promedio, entre 1991 y 2021, el contenido restrictivo de la regulación de la categoría transversal Otros fue de un 0,28 %.

Figura 3-13.
Proporción de expresiones vinculantes sobre el total de palabras de las regulaciones, categoría Otros



Fuente: elaboración propia con datos DNP-OMN.

Para mayor detalle correspondiente a las tendencias descritas de la proporción de expresiones vinculantes en el flujo total de palabras contenidas en las regulaciones sectoriales, se puede consultar la sección **"Anexos"** en donde, de forma separada, se grafican las tendencias de ambas variables para cada uno de los nueve sectores y la categoría *Otros*.

> **...Puede verse la gran heterogeneidad entre los valores tanto del flujo del total de palabras como del de expresiones vinculantes; esa característica implica que el flujo de regulación en los diferentes sectores no es una variable con una tendencia constante. Dicha variabilidad puede estar asociada a las dinámicas particulares que a lo largo del periodo 1991-2021 han requerido intervención regulatoria...**

Nuevas métricas REGCOL 3.0: expresiones vinculantes relevantes y medidas de complejidad de texto

Al considerar que la regulación es uno de los principales mecanismos a partir de los cuales se implementan las diferentes decisiones de política económica en un país, la calidad y la eficiencia de los actos administrativos puede estar asociada al desarrollo y crecimiento económico. Por consiguiente, se hace indispensable contar con medidas que, en primer lugar, contemplen la transversalidad y la relación entre regulaciones, y, en segundo lugar, que den cuenta de la complejidad de los actos administrativos expedidos, en especial si se tiene en cuenta que ello puede ser una aproximación sustancial a la calidad de los textos.

Según diferentes autores, los textos regulatorios usualmente no identifican exactamente el o los sectores hacia los cuales están direccionados e incluso pueden llegar a estar asociados a más de un sector en particular, por ende, se sugiere contar con medidas den cuenta de estas relaciones (Al-Ubaydli & McLaughlin, 2017; Baker *et al.*, 2016; Dawson & Seater, 2013). Por otro lado, de acuerdo con Alshner (2012) y Scheehan (2015) contar con métricas que permitan evaluar la legibilidad de las regulaciones facilita, mejora y acelera las reformas legales necesarias para un mayor acceso a la justicia, una menor carga regulatoria e incremento de la eficiencia.

Por consiguiente, esta fase del proyecto de REGCOL provee valor agregado en la medida en que permite construir métricas, que den cuenta de la asociación de la regulación con los diferentes sectores económicos y de la complejidad de los textos regulatorios, basadas en los documentos publicados en el *Diario Oficial* de la Imprenta Nacional de Colombia. En la versión actual del proyecto proporciona, por un lado, las expresiones vinculantes relevantes, que corresponden al ajuste en la medida de la restrictividad de los actos administrativos por la probabilidad de asociación con los distintos sectores económicos; y por otro lado, las métricas de cuentas condicionales relevantes, Dale-Chall y entropía de Shannon para medir y analizar el grado de complejidad de los actos administrativos en términos de entendimiento por grados de escolaridad. Así, REGCOL 3.0 busca contribuir con insumos para la toma de decisiones basadas en evidencia que favorezcan los procesos de construcción y ejecución de la Política de Mejora Regulatoria en el país.

4. 1 Expresiones vinculantes relevantes

En este numeral, se expondrán las proporciones de expresiones vinculantes relevantes que, a diferencia de las mostradas en la sección anterior están ajustadas por la probabilidad de asociación de cada acto administrativo de carácter sustancial con los sectores económicos, lo que implica un ajuste de la restrictividad por la subestimación o sobreestimación al no tener en cuenta anteriormente dicha probabilidad. Así, el ajuste de las proporciones por la probabilidad de asociación de cada acto administrativo con los distintos sectores económicos se calcula de la siguiente forma:

Las expresiones vinculantes relevantes al sector j en el año t son:

$$expre\_vinc\_rele_{j,t} = \sum_{i=1}^{n} (expre\_vinc_{i,t} * Pr_{i,j,t})$$

donde $Pr_{i,j,t}$ es la probabilidad de asociación del documento i con el sector j en el año t.

De forma equivalente, el total de palabras relevantes al sector j en el año t son:

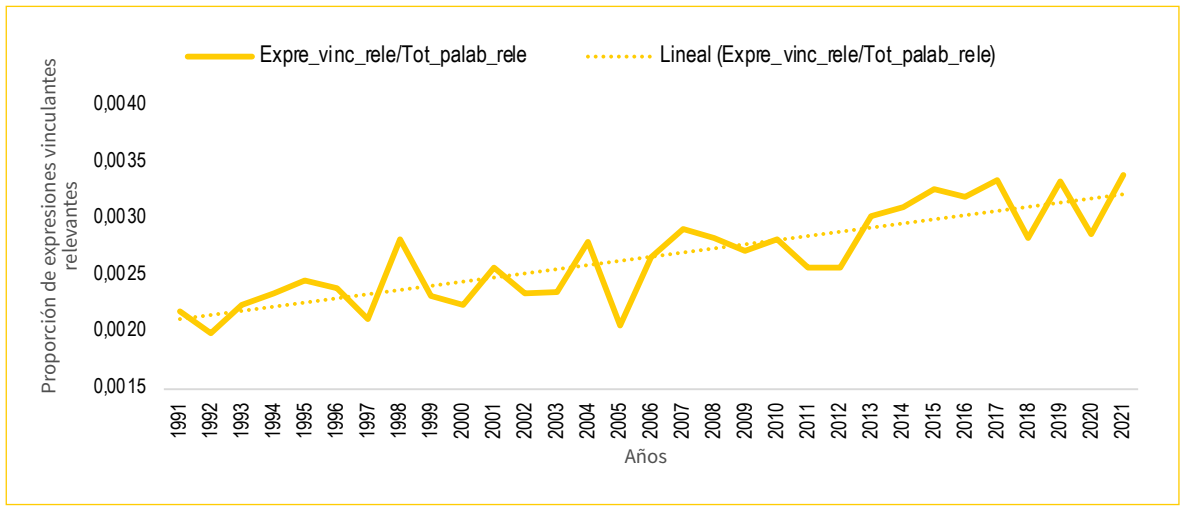
$$tot\_palab\_rele_{j,t} = \sum_{i=1}^{n} (tot\_palab_{i,t} * Pr_{i,j,t})$$

Dado esto, la proporción de expresiones vinculantes relevantes contenidas en las regulaciones de todos los sectores para el año t son:

$$\frac{expre\_vinc\_rele_t}{tot\_palab\_rele_t} = \frac{\sum_{j=1}^{J} expre\_vinc\_rele_{j,t}}{\sum_{j=1}^{J} tot\_palab\_rele_{j,t}}$$

En el inicio del análisis, se obtiene que la evolución de la proporción de expresiones vinculantes relevantes para los nueve sectores productivos sigue siendo creciente, pero se encontraba sobreestimada en 0,03 p. p., toda vez que al ajustarse la serie por la probabilidad de asociación el pico de 2007 se suaviza, al ubicar así el promedio general en el 0,26 % y no en el 0,29 % **(figura 4-1)**.

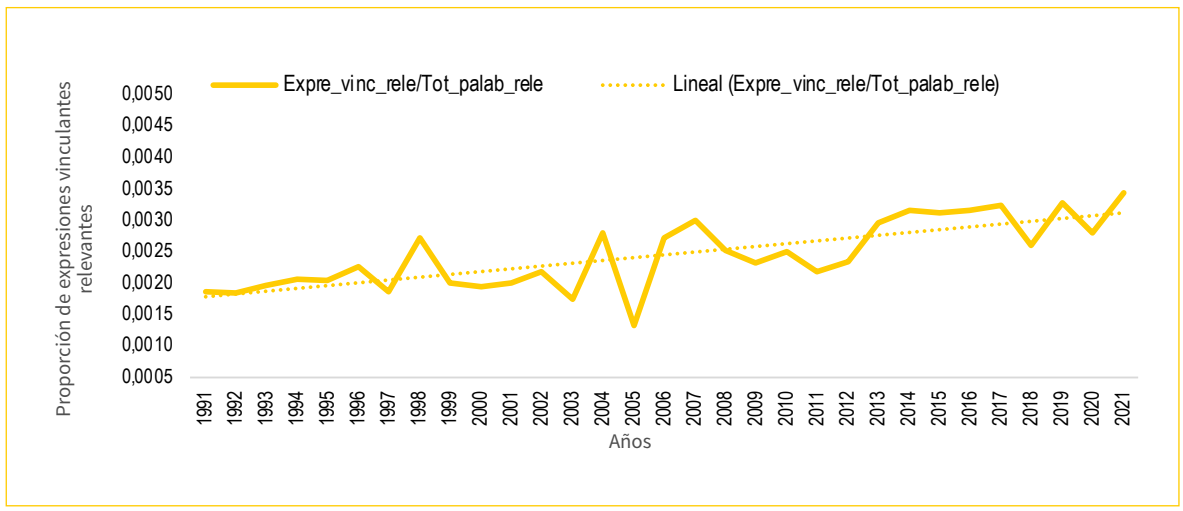
Figura 4-1.
Evolución de la proporción de expresiones vinculantes relevantes respecto al total de palabras relevantes de decretos y leyes que regulan nueve sectores productivos en Colombia, 1991-2021



Fuente: elaboración propia con datos OMN-DNP.

En cuanto al análisis por sectores, el sector *agropecuario, silvícola y pesquero* sigue manteniendo una tendencia creciente en todo el periodo de estudio, sin embargo, ya no muestra los dos picos de 2013 y 2019, ello indica que la restrictividad en estos dos periodos estaba sobreestimada. Por su parte, el promedio entre 1991 y 2021 fue mayor en 0,01 p. p., al pasar del 0,24 % al 0,25 % (**figura 4-2**).

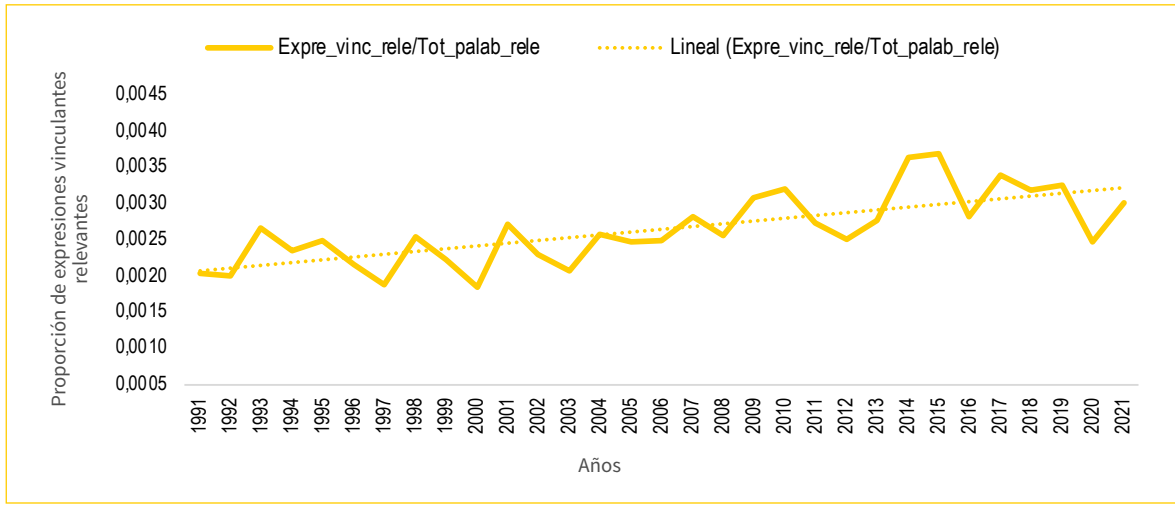
Figura 4-2.
Proporción de expresiones vinculantes relevantes sobre el total de palabras relevantes de las regulaciones, sector agropecuario, silvícola y pesquero



Fuente: elaboración propia con datos OMN-DNP.

Al analizar la proporción de expresiones vinculantes relevantes del sector *minería*, se demuestra una tendencia mucho más creciente al ajustarse por la probabilidad de asociación de los actos administrativos con dicho sector, toda vez que no se muestra el pico bajo de 1998, los repuntes de 2011, 2014 y 2017, y las caídas en 2018 y 2021 (**figura 4-3**), de modo que la evolución de la proporción de expresiones vinculantes estaba siendo subestimada en esos periodos. El promedio general pasó del 0,25 % al 0,26 %.

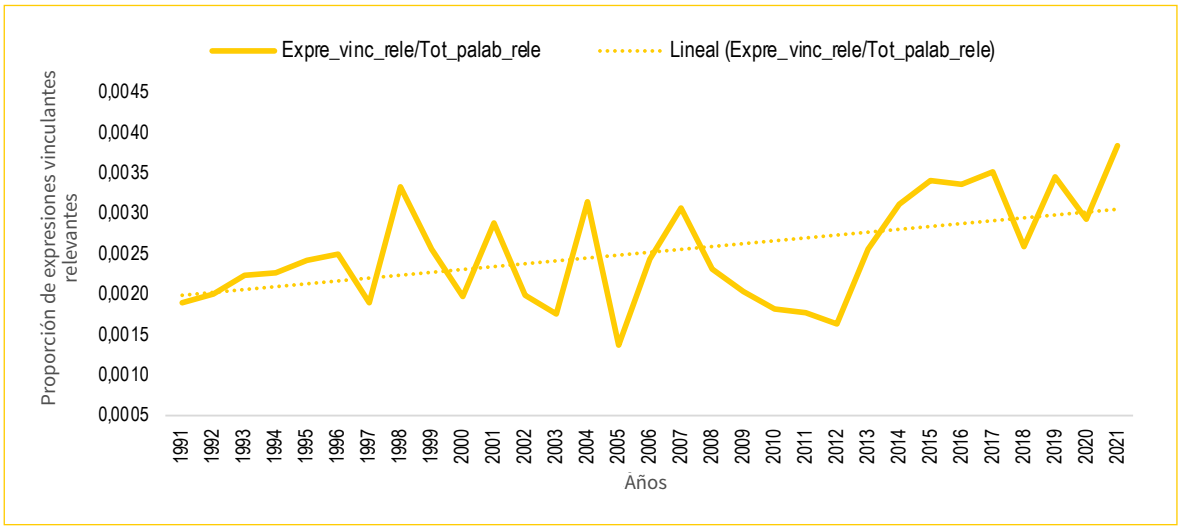
Figura 4-3.
Proporción de expresiones vinculantes sobre el total de palabras de las regulaciones, sector de minería



Fuente: elaboración propia con datos OMN-DNP.

Por su parte, la proporción de expresiones vinculantes relevantes del sector de *manufacturas* sigue mostrando una tendencia creciente durante todo el periodo de estudio. No obstante, la disminución hallada en 1992 ya no está presente al ajustar la serie por asociación de los actos administrativos con dicho sector; es decir, la proporción de palabras vinculantes sobre el total de palabras estaba subestimada (**figura 4-4**). En promedio, entre 1991 y 2021, el contenido restrictivo de la regulación del sector fue del 0,25 %, lo que implica que la evolución de la restrictividad en el sector estaba sobreestimada.

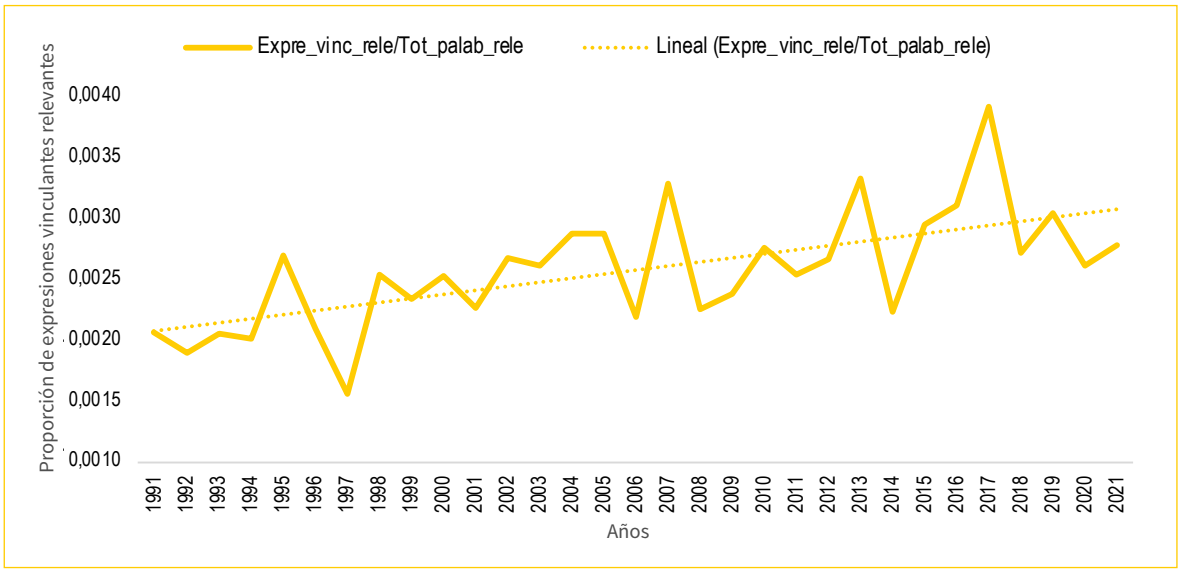
Figura 4-4.
Proporción de expresiones vinculantes sobre el total de palabras de las regulaciones, sector de manufacturas



Fuente: elaboración propia con datos OMN-DNP.

En cuanto a la proporción de expresiones vinculantes relevantes en el sector de *electricidad, gas y agua*, se ilustra que la tendencia tiene un crecimiento moderado al ajustarse por la probabilidad de asociación de los actos administrativos con el sector (**figura 4-5**). Las líneas se suavizan tanto en los picos altos como en los bajos, particularmente los que se ocurrieron en 2007 y 2015, y se obtuvo así un promedio de restrictividad en toda la serie del 0,26 %.

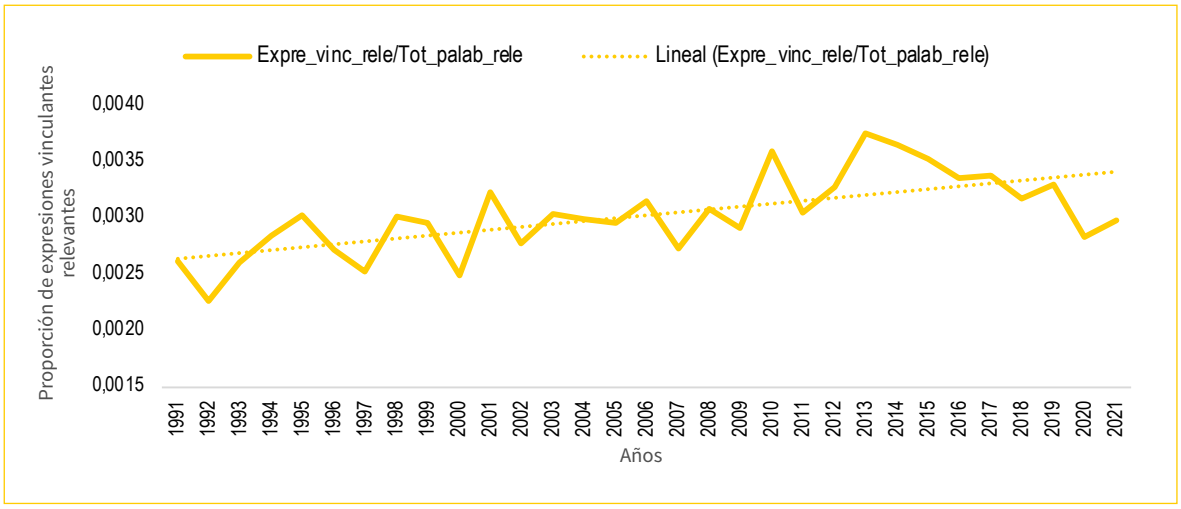
Figura 4-5.
Proporción de expresiones vinculantes sobre el total de palabras de las regulaciones, sector de electricidad, gas y agua



Fuente: elaboración propia con datos OMN-DNP.

En el sector de *construcción* cambió la tendencia decreciente a creciente en todo el periodo de estudio (figura 4-6), hecho que responde al hecho de que la proporción de 1992 pasó del 0,52 % al 0,28 %, lo cual implica que ese valor para dicho año estaba sobreestimado. Así mismo, el promedio en todo el periodo de estudio disminuyó 0,03 p. p. al ubicarse en el 0,30 %.

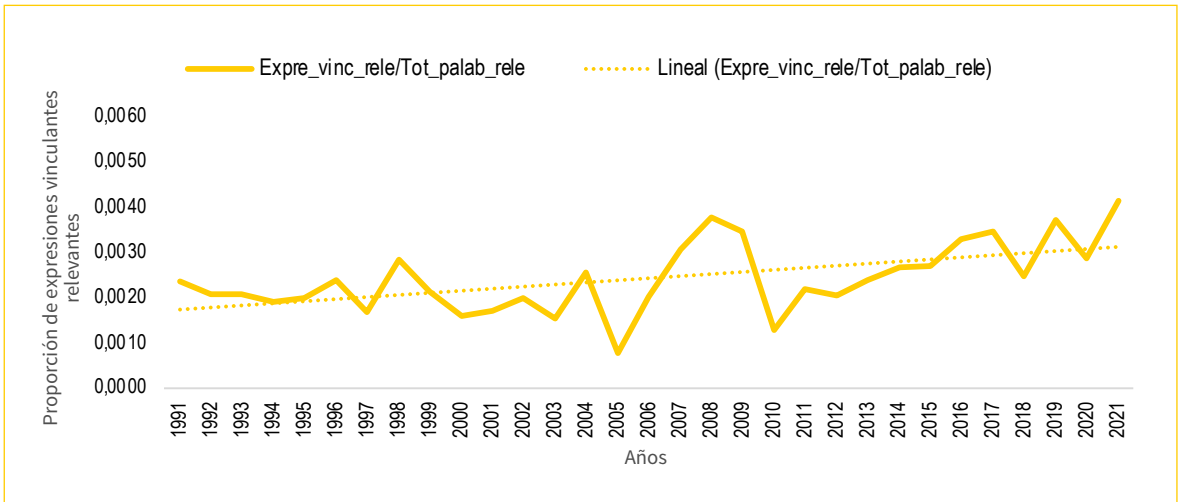
Figura 4-6.
Proporción de expresiones vinculantes sobre el total de palabras de las regulaciones, sector de construcción



Fuente: elaboración propia con datos OMN-DNP.

Al continuar con el sector de *comercio, restaurantes y hoteles*, al ajustarse la serie por la probabilidad de asociación de los actos administrativos con este sector, se puede observar que los cambios en la evolución de toda la serie fueron mínimos (**figura 4-7**). Lo cual indica que en este sector en particular no se estaba subestimado ni sobreestimación la proporción de expresiones vinculantes sobre el número total de palabras, toda vez que el promedio en su restrictividad siguió situándose en 0,26 %.

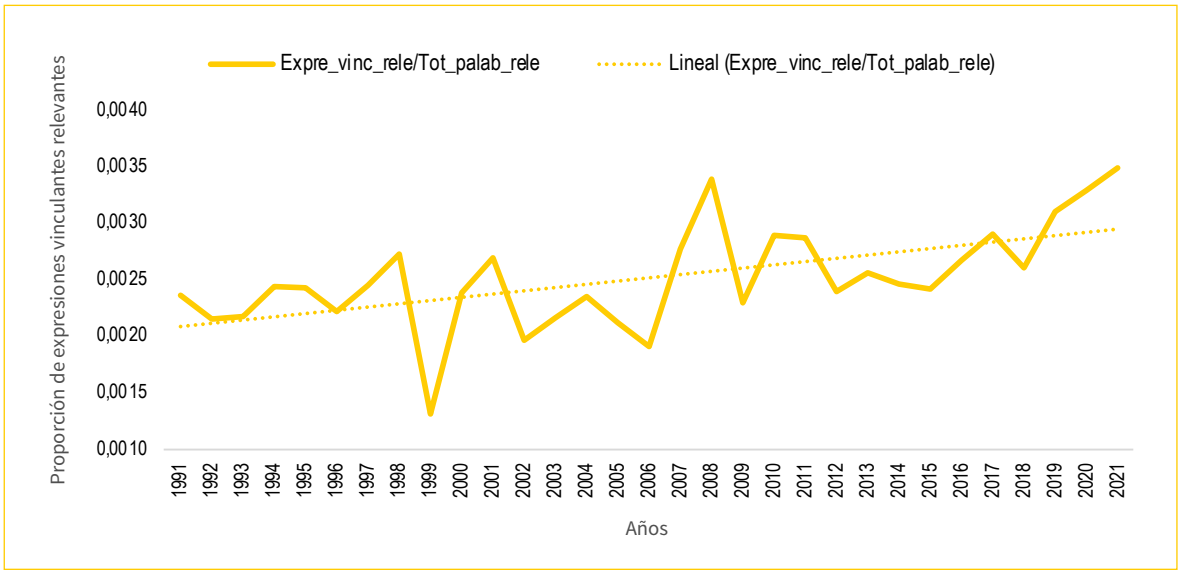
Figura 4-7.
Proporción de expresiones vinculantes sobre el total de palabras de las regulaciones, sector de comercio, restaurantes y hoteles



Fuente: elaboración propia con datos OMN-DNP.

Al analizar la evolución de la proporción de expresiones vinculantes en el sector de *transporte y comunicaciones*, hubo un cambio importante en el valor anteriormente registrado en 2016 al pasar del 0,84 % al 0,24 %; es decir, para este año en particular la cifra estaba sobreestimada en 0,6 p. p. Se obtiene así, un promedio en todo el periodo de análisis del 0,25 % (0,05 p. p. menos).

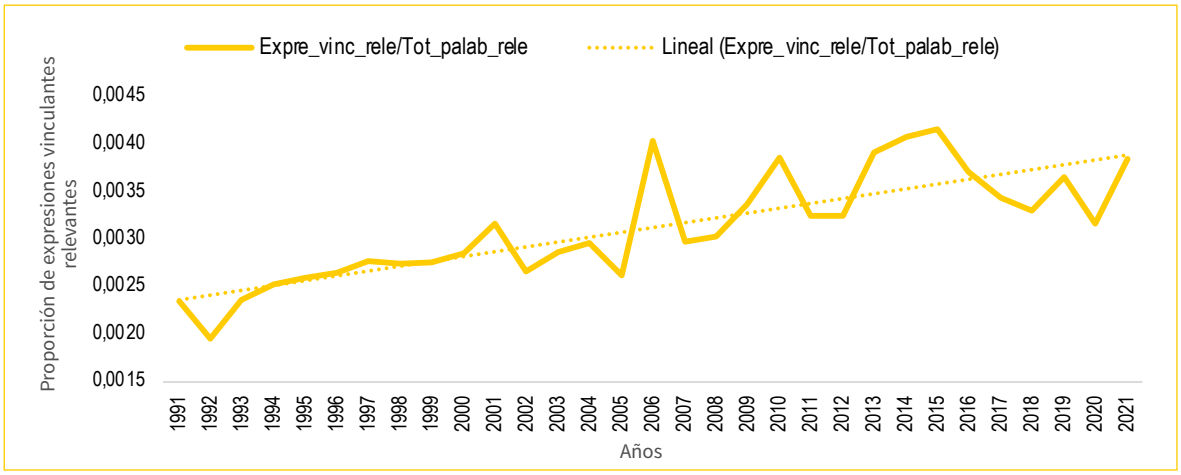
Figura 4-8.
Proporción de expresiones vinculantes sobre el total de palabras de las regulaciones, sector de transporte y comunicaciones



Fuente: elaboración propia con datos OMN-DNP.

En la **figura 4-9** se ilustra la razón de expresiones vinculantes relevantes del sector de *servicios financieros y empresariales*. Es notorio que al igual al sector de comercio, la evolución de la serie se mantiene muy parecida sin ajustar por probabilidad de asociación de los actos administrativos con dicho sector, sino que solo se percibe un pequeño cambio en el pico reportado en 2019. En promedio, entre 1991 y 2021, el contenido restrictivo de la regulación fue del 0,35 %.

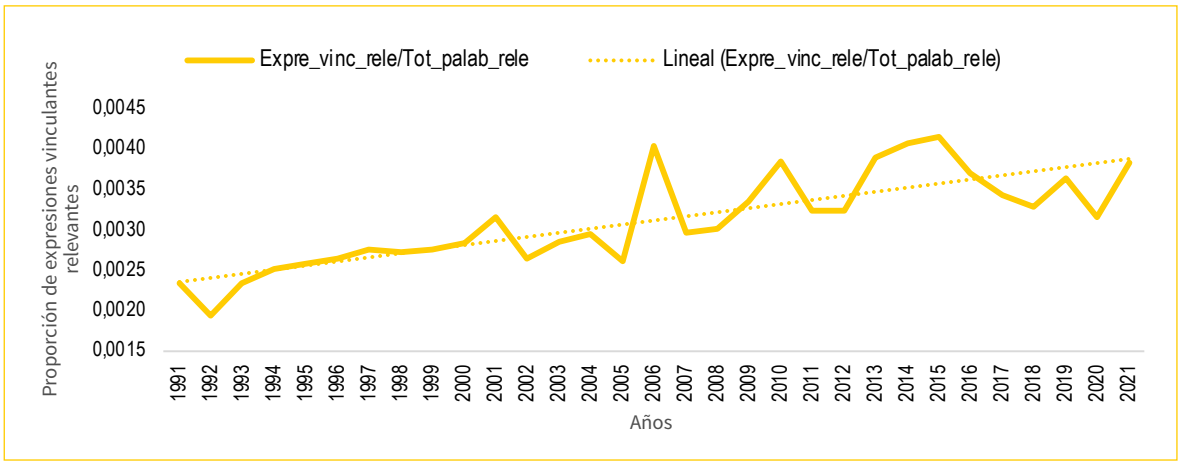
Figura 4-9.
Proporción de expresiones vinculantes sobre el total de palabras de las regulaciones, sector de servicios financieros y empresariales



Fuente: elaboración propia con datos OMN-DNP.

La dinámica en el tiempo de la proporción de expresiones vinculantes relevantes para el sector de *servicios comunales, sociales y personales* sí registra cambios en sus variaciones, toda vez que los picos hallados en 1995, 1998 y 2003 se suavizaron con al ajustarse por la probabilidad de asociación **(figura 4-10)**; es decir, ya no existe subestimación ni sobreestimación de las proporciones calculadas. Ahora el promedio general de la serie corresponde al 0,27 % —0,01 p. p. menos que sin ajustar por probabilidad—.

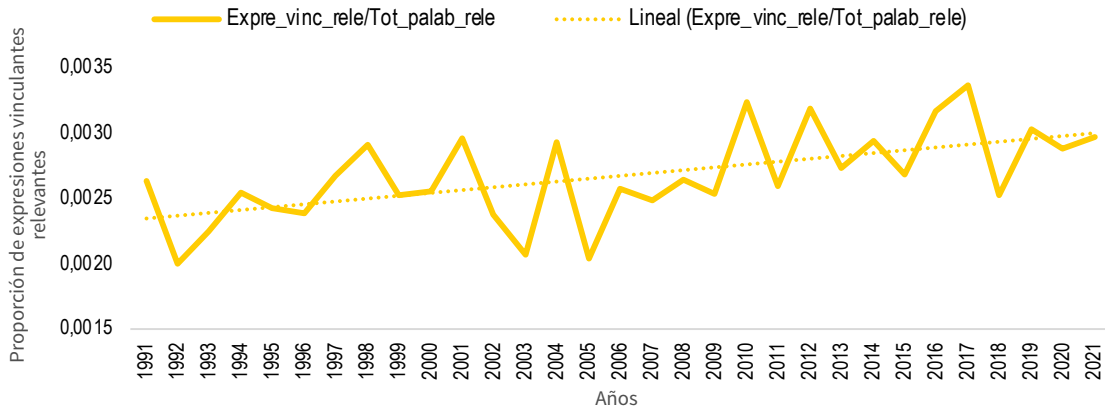
Figura 4-10.
Proporción de expresiones vinculantes sobre el total de palabras de las regulaciones, sector de servicios comunales, sociales y personales



Fuente: elaboración propia con datos OMN-DNP.

Finalmente, la tendencia de la proporción de palabras vinculantes relevantes de la categoría *Otros servicios* al igual que el sector de *comercio y servicios financieros* mantiene la misma tendencia y variaciones con el ajuste de la probabilidad de asociación (**figura 4-11**). Ello implica que no hubo ningún tipo de subestimación o sobreestimación con el contenido restrictivo en la categoría.

Figura 4-11.
Proporción de expresiones vinculantes sobre el total de palabras de las regulaciones, categoría Otros



Fuente: elaboración propia con datos OMN-DNP.

Para un mayor detalle en la comparación de la evolución de la proporción de expresiones vinculantes relevantes con y sin ajuste de la probabilidad de asociación de los actos administrativos a los sectores económicos, se puede consultar la sección de **"Anexos"** donde se presentan de forma paralela las tendencias de ambas series para cada uno de los nueve sectores y la categoría *Otros*.

4. 2 Métricas de complejidad de texto

Aquí se describen las nuevas métricas de complejidad de la regulación disponibles en la presente versión del proyecto REGCOL, y que son referencias internacionales para este tipo de ejercicios.

4.2.1 Métrica de cuentas condicionales relevantes

El primer enfoque metodológico hace uso de una función matemática que tiene como objetivo cuantificar todas las posibles asociaciones condicionales que se encuentran en un texto para después estimar la cantidad total de condicionales en un texto. Intuitivamente esto representa la cantidad de ramificaciones lógicas que separan las diferentes proposiciones e ideas (bloques de ideas) identificados dentro de un texto regulatorio.

El algoritmo con el cual se estructura esta métrica se resume en la siguiente secuencia de pasos:

1. Definir el listado de condiciones, conectores lógicos y expresiones condicionales del lenguaje que serán empleados en el análisis. Ese listado lo construyeron expertos temáticos de la Subdirección de Gobierno y Asuntos Internacionales (SGAI) y se cargó al proceso de cuantificación de la métrica.

2. Estructurar cada texto regulatorio mediante un proceso de vectorización en el cual se crea una matriz de características que asigna una columna por palabra mientras que cada fila corresponde a una revisión. En esta versión del proyecto se empleó el método denominado *bolsa de palabras* como método de vectorización del corpus de textos regulatorios.

3. Recorrer la lista de palabras condicionales en la columna de la base de datos que contiene las palabras del texto únicas y, si existe coincidencia, se añade la frecuencia de esa palabra en un objeto tipo lista.

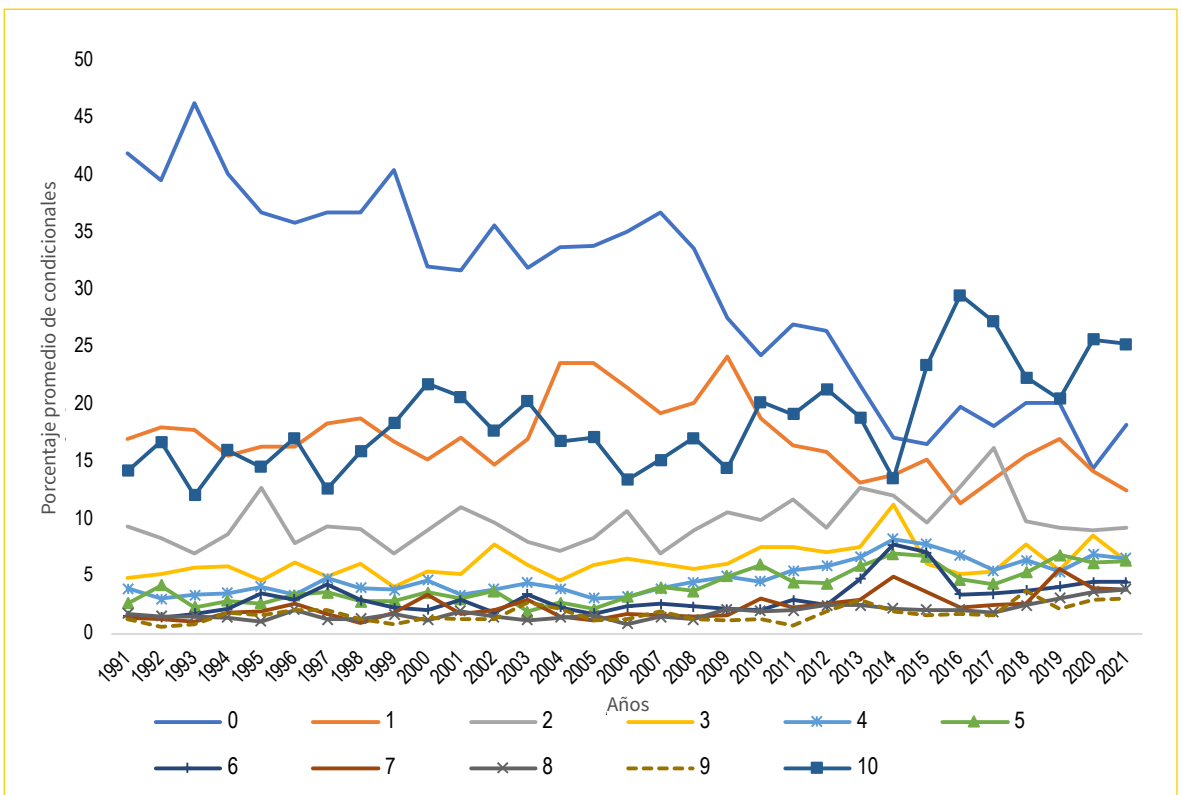
4. Generalizar la base de datos consolidada en el paso anterior mediante la función de pérdida considerada para mapear la estructura de cada texto regulatorio. Dicha función consiste en el total de condicionales identificados ajustado por la probabilidad de asociación de cada acto administrativo con los distintos sectores económicos por año de análisis $(P_{i,j,k})$; es decir, la fórmula matemática para estimar la métrica de cuentas condicionales en el j-ésimo sector económico para el año k está dado por:

$$CC_{i,j,k} = \sum_{i=1}^{n_i} C_{i,j,k} \, P_{i,j,k}$$

Cada una de estas frecuencias se suma y produce como resultado el indicador de cuentas condicionales para un solo texto.

La lista de condiciones se basa en expresiones condicionales definidas que permiten separar o ramificar así ideas en bloques. A continuación se ofrecen los resultados obtenidos por la métrica de cuentas condicionales para los textos regulatorios no sustanciales (0) y los sustanciales caracterizados en los nueve sectores productivos y en la clasificación transversal *Otros* (10). En la **figura 4-12** se presenta la evolución temporal del porcentaje promedio de condicionales en los textos regulatorios contenidas en los decretos y leyes durante el periodo de interés. Resulta importante destacar que en tanto pasa el tiempo el porcentaje promedio de condicionales identificados en las regulaciones no sustanciales ha disminuido considerablemente, mientras que el porcentaje promedio de condicionales en el sector transversal aumenta. En los demás sectores económicos no se observa un patrón específico en la métrica.

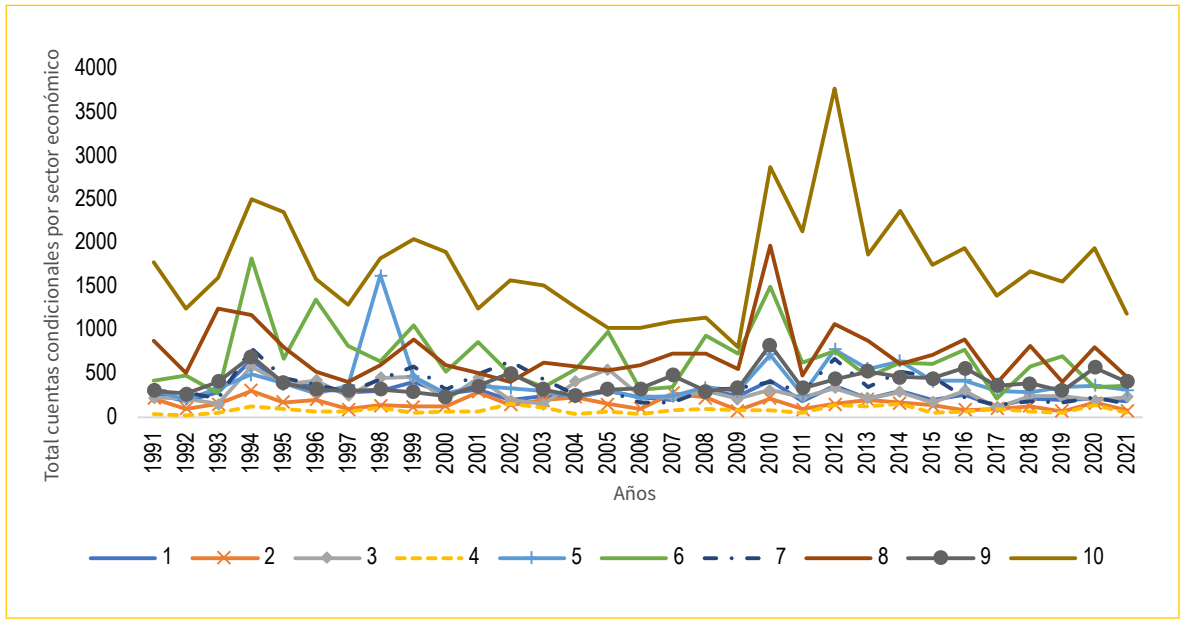
Figura 4-12.
Evolución del porcentaje promedio de condicionales en el corpus regulatorio sustancial y no sustancial de Colombia, 1991-2021



Fuente: elaboración propia con datos OMN-DNP.

El análisis continuará con la revisión de los resultados de la métrica al quitar los textos clasificados como no sustanciales (0). En la **figura 4-13** describe la evolución temporal del número de cuentas condicionales en los textos regulatorios contenidas en los decretos y leyes durante el periodo de interés. La figura muestra que el mayor número de cuentas condicionales se da en el sector transversal *Otros* (10) a lo largo de todo el lapso determinado; por ello, se considera ideal continuar el análisis sin este sector. En el resto de los sectores no se observa una tendencia específica de interés particular; se puede resaltar un pico a lo largo de los sectores en el año 2010.

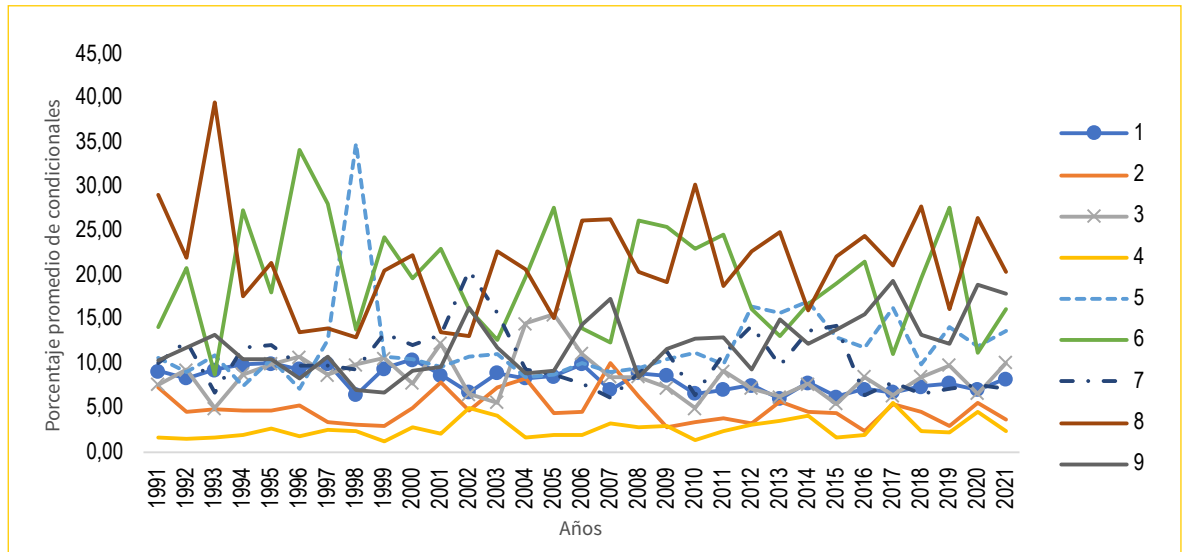
Figura 4-13.
Evolución de las cuentas condicionales en el corpus regulatorio sustancial de Colombia, 1991-2021



Fuente: elaboración propia con datos OMN-DNP.

Finalmente, en la **figura 4-14** señala la evolución temporal del porcentaje promedio de cuentas condicionales en los textos regulatorios contenidas en los decretos y leyes de los nueve sectores económicos de Colombia. En las líneas puede observarse que los sectores de *servicios financieros y empresariales*, junto con el sector de *comercio, restaurantes y hoteles* son los que tienen el mayor porcentaje promedio de cuentas condicionales. Por el contrario, los sectores con el menor porcentaje promedio de cuentas condicionales son los sectores de *electricidad, gas y agua,* y de *minería* a lo largo de todo el periodo de interés. Adicional a lo anterior, no se observa ninguna tendencia creciente o decreciente de la métrica a lo largo del tiempo; sin embargo, sobresale un gran pico en el año 1998 en el sector de *construcción*.

Figura 4-14.
Evolución del porcentaje promedio de condicionales en el corpus regulatorio sustancial de los sectores económicos en Colombia, 1991-2021



Fuente: elaboración propia con datos OMN-DNP.

4.2.2 Métrica Dale-Chall

Es la métrica en la cual se usa un conteo de palabras "difíciles" en lugar de emplear la longitud de las palabras para evaluar la complejidad de un texto. Para tal fin, se utiliza un listado de 3.000 palabras familiares y comunes en el lenguaje con el objetivo de estimar el puntaje de facilidad de lectura. El algoritmo hace uso de las siguientes ecuaciones para la estimación de la métrica:

$$PI = 0.1519 * PPD + 0.0496 * LPO$$

Siendo

PI: puntaje inicial.
PPD: porcentaje de palabras difíciles en el texto.
LPO: longitud promedio de una oración en número de palabras.

Según la metodología, si el $PPD > 5\ \%$, entonces,

$$PA = PI + 3.6365$$

En cualquier otro caso, $PA = PI$, donde PA es el puntaje de la métrica Dale-Chall, la cual relaciona el grado de comprensión de lectura que tiene lector a partir del 4.° grado como nivel de referencia. En la **tabla 4-1** se enlista la interpretación de los resultados obtenidos en los textos regulatorios en análisis, según el nivel educativo de la persona capaz de entender el texto.

Tabla 4-1

Puntaje de la métrica Dale-Chall y su relación con el nivel educativo de la persona capaz de entender el texto

Puntaje Dale-Chall	Nivel educativo
De 4,9 o menos	4.° grado o menos
De 5,0 a 5,9	Grados 5.° y 6.°
De 6,0 a 6,9	Grados 7.° y 8.°
De 7,0 a 7,9	Grados 9.° y 10.°
De 8,0 a 8,9	Grados 11.°y 12.°
De 9,0 a 9,9	Universitario
De 10 o más	Posgrado

Fuente: Traducción del artículo de Chall & Dale (2000).

El proceso de estimación de métrica Dale-Chall se dividió en dos etapas las cuales serán descritas a continuación.

1. Estimación de proporción de palabras “difíciles”:

- **a.** Depurar y vectorizar los textos regulatorios.
- **b.** Calcular la cantidad de palabras del texto.
- **c.** Realizar el conteo de las palabras de uso común dada en la lista de Dale-Chall.
- **d.** Calcular la proporción de palabras en el texto que no pertenecen a la lista de Dale-Chall.

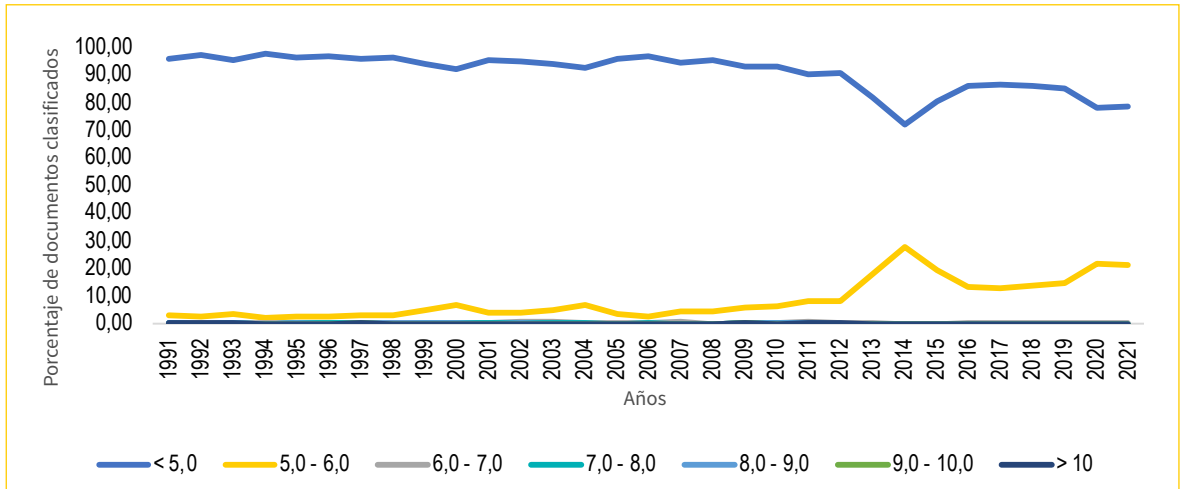
2. Estimación de longitud de cadena de caracteres:

- **a.** Dividir las palabras por letra y calcular la longitud de la lista por texto regulatorio.
- **b.** Estimar la cantidad de caracteres por texto.
- **c.** Calcular el cociente entre la cantidad de caracteres del texto y el número de palabras que lo forman, la cual se define como una medida aproximada del promedio de caracteres por palabra.

Los resultados de indicador de comprensión de Dale-Chall (Dale & Chall, 1948) como predictor de la dificultad de lectura de los textos regulatorios conforman el tema de este y los siguientes apartados. En la **figura 4-15** plantea la evolución temporal del porcentaje de textos regulatorios (decretos y leyes) con puntajes de 4,9 o menos, de 5,0 a 5,9, de 6,0 a 6,9, de 7,0 a 7,9, de 8,0 a 8,9, de 9,0 a 9,9 y de 10 o más, durante el periodo de interés. En la representación gráfica puede observarse que la gran mayoría de textos son de muy fácil comprensión; es decir, con puntajes de 4,9 o menos (4 grado o menos) y de 5,0 a 5,9 (grados 5.° y 6.°). Resulta importante destacar que a medida

que pasa el tiempo el porcentaje de textos regulatorios (decretos y leyes) con puntajes de 4,9 o menos ha disminuido regularmente a partir del 2012 de forma similar a como ha aumentado el porcentaje de documentos con una dificultad ligeramente mayor con puntaje de 5,0 a 5,9 (grados 5.° y 6.°). El porcentaje de documentos con más dificultad de nivel universitario o mayor durante todo el periodo de tiempo analizado es muy bajo.

Figura 4-15.
Puntaje de la métrica Dale-Chall en el corpus regulatorio sustancial de los sectores económicos en Colombia, 1991-2021



Fuente: elaboración propia con datos OMN-DNP.

4.2.3 Métrica entropía de Shannon

Esta métrica crea un modelo matemático que permite cuantificar y descifrar un sistema de comunicación por medio de entidades de probabilidad que miden la dificultad promedio de un texto a partir de la probabilidad de ocurrencia de cada uno de los eventos (Ellerman, 2021). En el procesamiento de lenguaje natural la métrica se define como la cantidad de información promedio que contienen los símbolos usados en un texto regulatorio (Gray, 2011). La métrica de interés se obtiene a partir de un algoritmo donde el corpus de textos se define como el conjunto de acontecimientos o eventos los cuales ocurren uno a la vez, ellos conforman el sistema total que puede contener un número *n* de eventos $S = \{E_1, E_2, \ldots, E_n\}$.

En el procesamiento de lenguaje natural la entropía de Shannon considera como sistema de símbolos a las palabras de los textos regulatorios en el cual los símbolos con menor probabilidad son los que aportan más información. En tal sentido, las palabras ampliamente utilizadas como “que”, “el”, “a” aportan poca información, mientras que las menos frecuentes como “plurivalente”, “distopía”, “ucronía” aportan más. La interpretación de la métrica indica que si en un texto se borra una palabra

muy utilizada la interpretación no se afectará ampliamente; situación que cambiará de manera notoria si se borra una palabra de difícil comprensión. Cuando todos los símbolos son igualmente probables, todos aportan información relevante y la entropía es máxima. El proceso de estimación de la métrica se resume en la siguiente secuencia de pasos:

1. Estructurar cada texto regulatorio mediante un proceso de vectorización en el cual se crea una matriz de características que asigna una columna por palabra, mientras que cada fila corresponde a una revisión. En esta versión del proyecto se empleó el método denominado bolsa de palabras como mecanismo de vectorización del corpus de textos regulatorios.

2. Estimar la frecuencia de cada palabra y expresión del texto depurado.

3. Calcular la probabilidad de aparición de cada palabra en el texto regulatorio (p_i).

4. Calcular el logaritmo en base 2 de la probabilidad de aparición de cada palabra en el texto regulatorio (p_i) y el producto entre esa cantidad respecto a su probabilidad.

5. Calcular la entropía del texto. $P_i * \log_2 P_i$

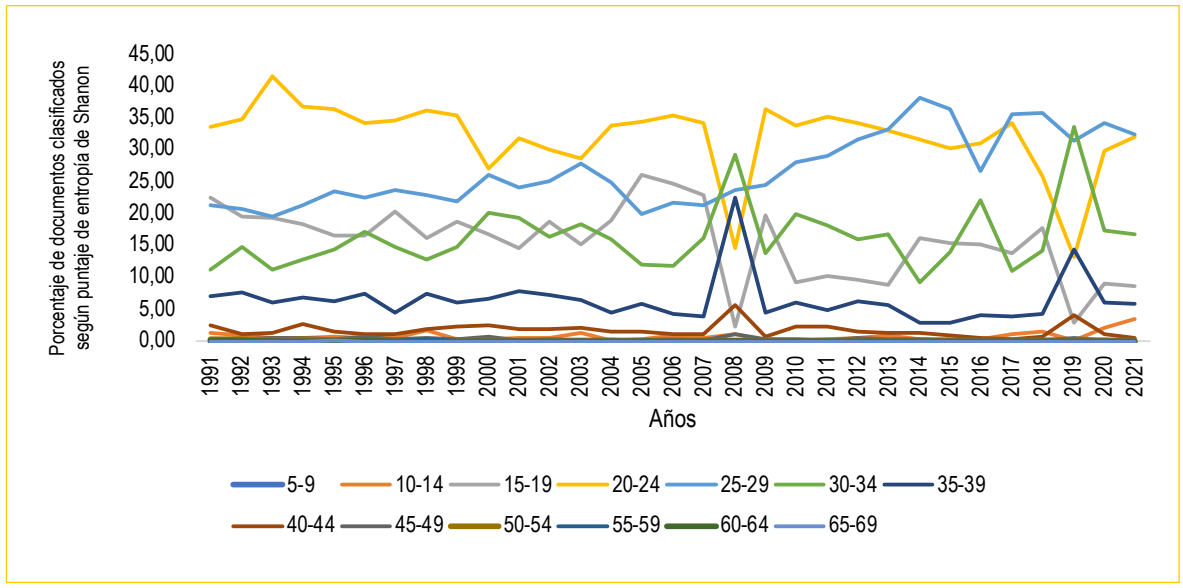
$$H = \sum_{i=1}^{n} P_i * \log_2 P_i$$

Es fundamental aclarar que esta métrica tiene una interpretación diferente a la métrica de Dale-Chall, ya que a mayor valor de la entropía de Shannon menos difícil resulta la comprensión de un texto. Ello se debe a que si la probabilidad de que una palabra aparezca en el texto es muy baja podría esperarse que fuesen términos técnicos muy específicos; es decir, de uso poco común; por lo tanto, un texto muy complejo de entender; esto significa que si tiene palabras muy técnicas dará como resultado un valor bajo de la entropía de Shannon (Mateos, 2016).

Ahora se presentan los resultados de la entropía de Shannon como predictor de la dificultad de lectura de los textos regulatorios. La figura 4-16 registra la evolución temporal del porcentaje de textos regulatorios (decretos y leyes) con puntajes de entropía de Shannon entre 5 y 9, 10 y 14, 15 y 19, 20 y 24, 25 y 29, 30 y 34, 35 y 39, 40

y 44, 45 y 49, 50 y 54, 55 y 59, 60 y 64, 65 y 69, durante el periodo de interés. Puede observarse que la gran mayoría de textos son de dificultad medio-alta; es decir, con puntajes de entropía de Shannon de 15 a 19, 20 a 24 y 25 a 29; también se tiene un porcentaje moderado la métrica que fluctúa de 30 a 34 y de 35 a 39. El porcentaje de documentos con un alto nivel de dificultad (nivel universitario o mayor) durante todo el periodo de tiempo analizado es bajo (entropía de Shannon entre 5 y 9, 10 y 14).

Figura 4-16.
Porcentaje de textos regulatorios sustanciales en los sectores económicos en Colombia (1991-2021) clasificados por su entropía de Shannon



Fuente: elaboración propia con datos OMN-DNP.

“...Los resultados de indicador de comprensión de Dale-Chall indican que la gran mayoría de textos son de muy fácil comprensión; es decir, con puntajes de 4,9 o menos (4 grado o menos) y de 5,0 a 5,9 (grados 5.° y 6.°). Sin embargo, a partir de 2012, esta proporción ha disminuido indicando que los textos se están volviendo en promedio más complejos.

En contraste, los resultados del indicador de Entropía de Shanon indican que la gran mayoría de textos son de dificultad medio-alta; es decir, con puntajes de entropía de Shannon de 15 a 19, 20 a 24 y 25 a 29...”

Conclusiones y pasos por seguir en el proyecto REGCOL

El proyecto REGCOL en su versión 3.0 presenta resultados actualizados al 2021 de las medidas de expresiones vinculantes y conteo total de palabras de tal forma que se da continuidad a los resultados obtenidos en las versiones anteriores 1.0 y 2.0. Adicionalmente, se formulan nuevas medidas de restrictividad usando como ponderador las probabilidades de asociación de cada texto regulatorio a los diferentes sectores económicos. La nueva aproximación recoge en mejor medida la característica transversal de las regulaciones y se considera como un notorio avance del proyecto en la construcción de medidas que den cuenta de las características intrínsecas de las regulaciones. En tal sentido, se avanzó también en la construcción e implementación de medidas de complejidad de textos y se abre la línea de estudio tanto de dichas características como de sus posibles implicaciones en el entendimiento y, por ende, el cumplimiento de las regulaciones.

Las métricas de expresiones vinculantes y palabras relevantes utilizadas en REGCOL 3.0 ofrecen una aproximación más precisa de la transversalidad de las regulaciones en diferentes sectores. Estas nuevas medidas confirman las tendencias observadas en mediciones anteriores, pero también suavizan las observaciones al considerar ajustes de restrictividad y ponderación de la probabilidad de asociación de cada texto regulatorio con los sectores económicos correspondientes. Con esto se tiene que, durante el periodo de análisis, la relación entre las expresiones vinculantes relevantes y el total de palabras relevantes contenidas en la regulación ha mantenido una tendencia creciente donde, las expresiones vinculantes relevantes representan en promedio el 0,26% del total de las palabras contenidas en los textos regulatorios.

Por su parte, de los resultados de las métricas de complejidad se deduce que, durante el periodo de análisis, los textos regulatorios publicados en el *Diario Oficial* de la Imprenta Nacional de Colombia tienden a tomar valores bajos en la escala de las métricas de complejidad de textos utilizadas en este estudio, lo cual permite afirmar que la mayoría de estos textos son de muy fácil comprensión respecto a la dimensión del uso de palabras comunes en el lenguaje. Por lo contrario, respecto a la dimensión de incertidumbre de la información tales textos se clasifican como de complejidad media-alta. Esa dualidad sirve de insumo para un análisis más riguroso y específico de la complejidad de textos regulatorios en Colombia que aborde el problema desde un enfoque multidimensional que desglose el problema en sus diferentes aspectos primarios en una próxima versión del proyecto REGCOL.

Los resultados del proyecto REGCOL ofrecen múltiples beneficios para funcionarios públicos, académicos y otros actores interesados en contribuir al desarrollo de la Política de Mejora Normativa en Colombia y mejorar la calidad de las regulaciones del país en general. Las expresiones vinculantes relevantes permiten evaluar el impacto de las regulaciones en diversos sectores. Al cuantificar el grado de restrictividad u obligatoriedad de cumplimiento implícito, se logra comprender mejor la carga regulatoria impuesta a los actores regulados y evaluar los posibles efectos económicos. Estos datos son valiosos para las evaluaciones de impacto regulatorio. La comparabilidad temporal de estas medidas facilita la identificación de los posibles efectos de la implementación de buenas prácticas, lo cual contribuye a la toma de decisiones basada en evidencia y a las reformas regulatorias. Por último, los investigadores pueden utilizar las medidas de restrictividad y complejidad para investigar la relación entre los entornos regulatorios y diversos resultados socioeconómicos. Estos indicadores proporcionan un marco estandarizado para estudiar los efectos de las regulaciones en la actividad empresarial, la innovación, la competencia y el rendimiento económico en general.

El Observatorio de Mejora Normativa seguirá en la tarea de actualización, mejora y aprovechamiento de la información disponible del proyecto REGCOL. Por ello, se proponen como líneas de trabajo las siguientes:

- Con el fin de optimizar los rendimientos de clasificación se recomienda agregar más normativas clasificadas manualmente en los sectores económicos a la base de datos requerida por el proceso de entrenamiento; especialmente en aquellos con menos documentos: minería y extracción; electricidad, gas y suministros de agua; construcción y transporte. Además, se explorarán nuevas metodologías de clasificación complementarias como el modelamiento de tópicos que posiblemente aportarán al proyecto REGCOL.

- Se explorará la ampliación de sectores disponibles de clasificación permitiendo una mayor desagregación de la información hacia sectores más específicos.

- Se seguirá trabajando en mejorar la implementación de las medidas de complejidad y claridad de textos y se hará conexión con la estrategia de lenguaje claro del Estado colombiano.

Bibliografía

Aggarwal, C. C. (2018). *Machine Learning for Text* (1st ed.). Springer.

Alschner, W. (2021). Validating Readability and Complexity Metrics: A New Dataset of Before-and-After Laws. SSRN Electronic Journal. https://doi.org/10.2139/ssrn.3812595

Al - Ubaydli, O., & McLaughlin, P. A. (2017). *RegData: A numerical database on industry-specific regulations for all United States industries and federal regulations, 1997–2012.* Regulation & Governance, 11(1), 109–123. https://doi.org/10.1111/rego.12107

Baker, S. R., Bloom, N., & Davis, S. J. (2016). Measuring Economic Policy Uncertainty*. *The Quarterly Journal of Economics, 131*(4), 1593–1636. https://doi.org/10.1093/qje/qjw024

Chall, J. S., & Dale, E. (2000). *Readability Revisted: The New Dale-Chall Readability Formula* (1st ed.). Brookline Books/Lumen Editions.

Dale, E., & Chall, J. S. (1948). *A Formula for Predicting Readability.* Educational Research Bulletin, 27, 37–54.

Dawson, J. W., & Seater, J. J. (2013). Federal regulation and aggregate economic growth. *Journal of Economic Growth*, 18(2), 137–177. https://doi.org/10.1007/s10887-013-9088-y

Ellerman, D. (2021). *New Foundations for Information Theory: Logical Entropy and Shannon* Entropy (1st ed.). Springer.

Evans, J. A., & Aceves, P. (2016). Machine Translation: Mining Text for Social Theory. *Annual Review of Sociology,* 42(1), 21–50. https://doi.org/10.1146/annurev-soc-081715-074206

Gray, R. M. (2011). *Entropy and Information Theory* (2nd ed.). Springer.

Grimmer, J., Roberts, M. E., & Stewart, B. M. (2022). *Text as Data: A New Framework for Machine Learning and the Social Sciences* (1st ed.). Princeton University Press.

Hvitfeldt, E., & Silge, J. (2021). *Supervised Machine Learning for Text Analysis in R* (1st ed.). Chapman & Hall/ CRC.

Mateos, D. M. (2016). *Medidas de complejidad y de información como herramientas para el análisis de series temporales.* Universidad Nacional de Córdoba.

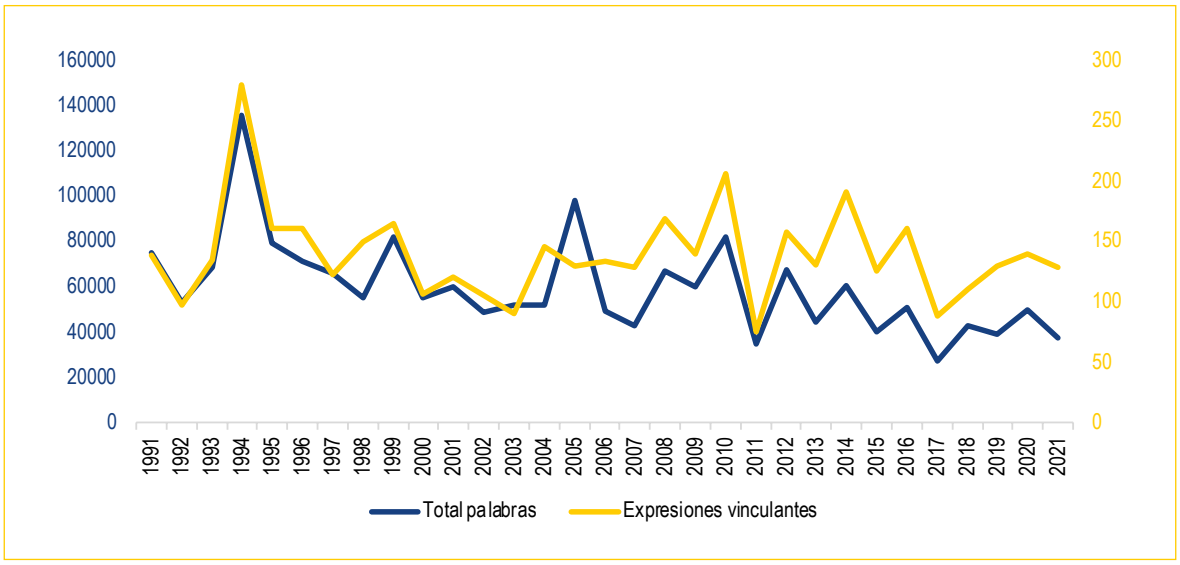
Observatorio de Mejora Normativa. (2021). *REGCOL 2.0: ampliación del periodo disponible (1991-2019) y mejoramiento del modelo de clasificación sectorial*.DNP.

Sheehan, K. M. (2015). Aligning TextEvaluator ® Scores With the Accelerated Text Complexity Guidelines Specified in the Common Core State Standards. *ETS Research Report Series, 2015*(2), 1–20. https://doi.org/10.1002/ets2.12068

Anexos

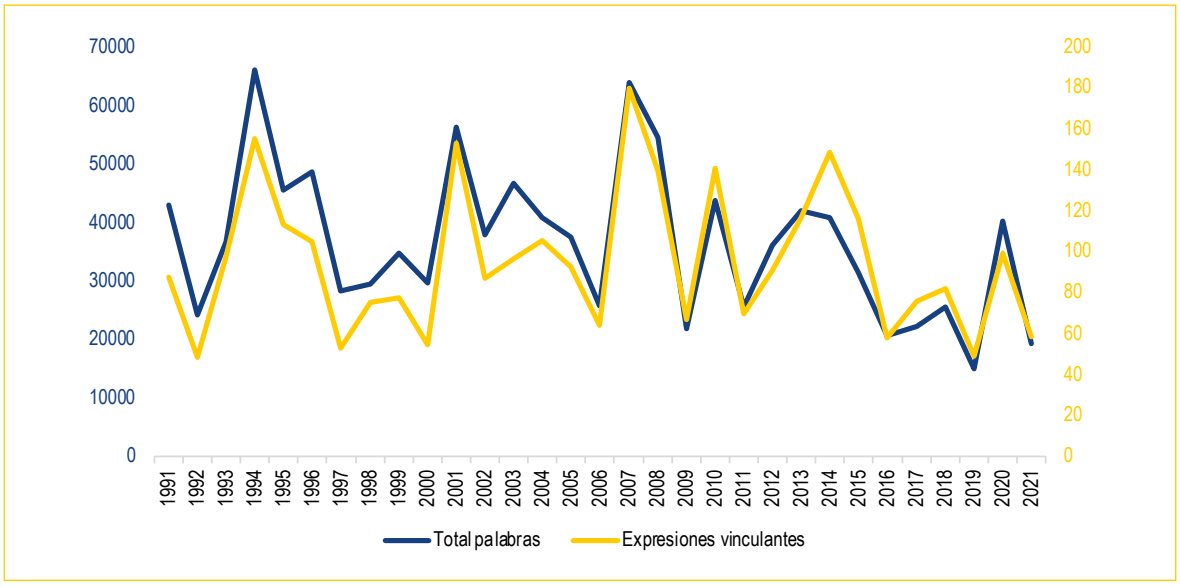
A. Expresiones vinculantes y total de palabras de las regulaciones, por sector

Figura 7A -1.
Expresiones vinculantes y total de palabras de las regulaciones, sector agropecuario, silvícola y pesquero



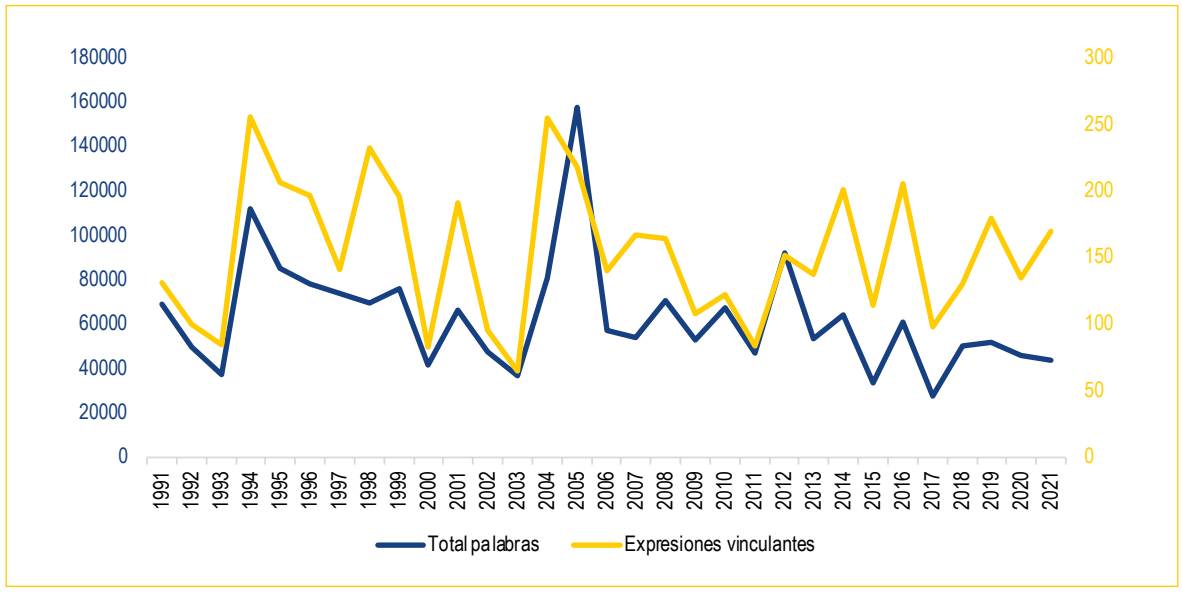
Fuente: elaboración propia con datos OMN-DNP.

Figura 7A -2.
Expresiones vinculantes y total de palabras de las regulaciones, sector de minería



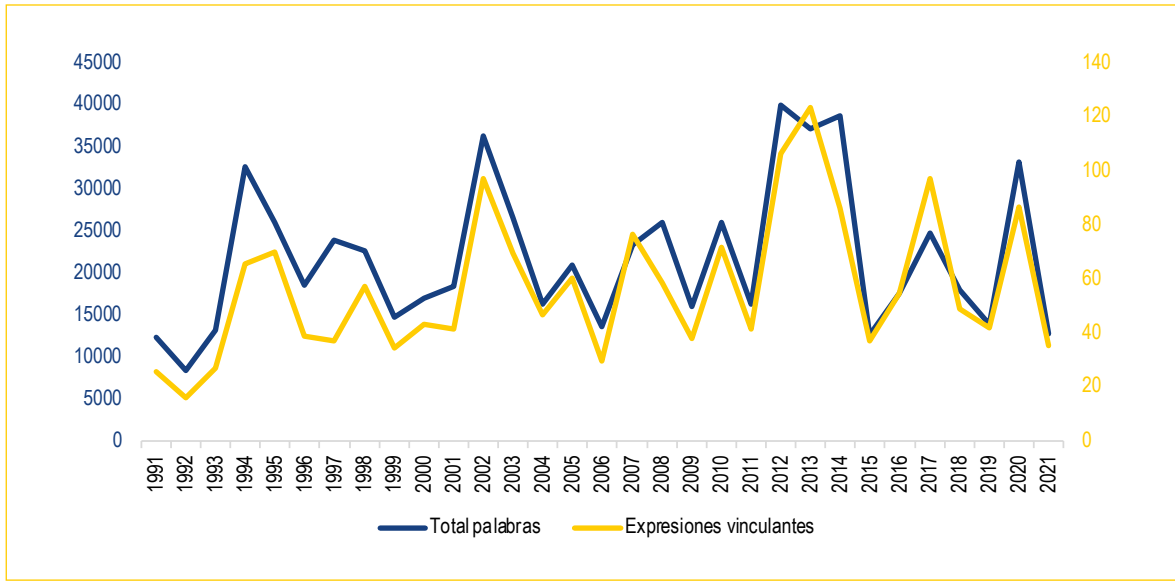
Fuente: elaboración propia con datos OMN-DNP.

Figura 7A -3.
Expresiones vinculantes y total de palabras de las regulaciones, sector de manufactura



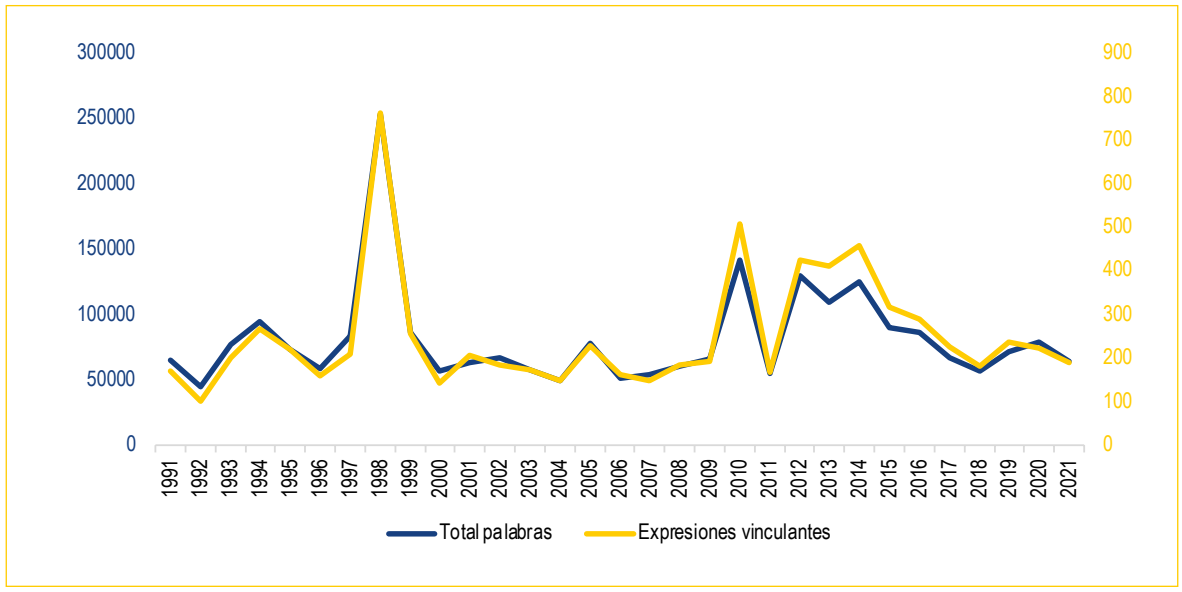
Fuente: elaboración propia con datos OMN-DNP.

Figura 7A -4.
Expresiones vinculantes y total de palabras de las regulaciones, sector de electricidad, gas y agua



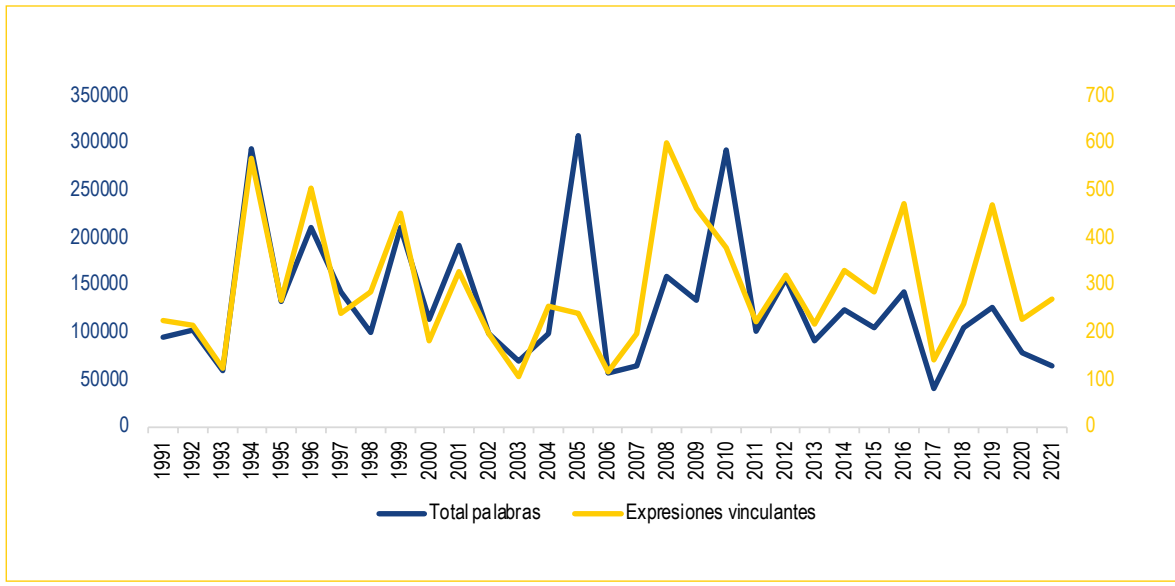
Fuente: elaboración propia con datos OMN-DNP.

Figura 7A -5.
Expresiones vinculantes y total de palabras de las regulaciones, sector de construcción



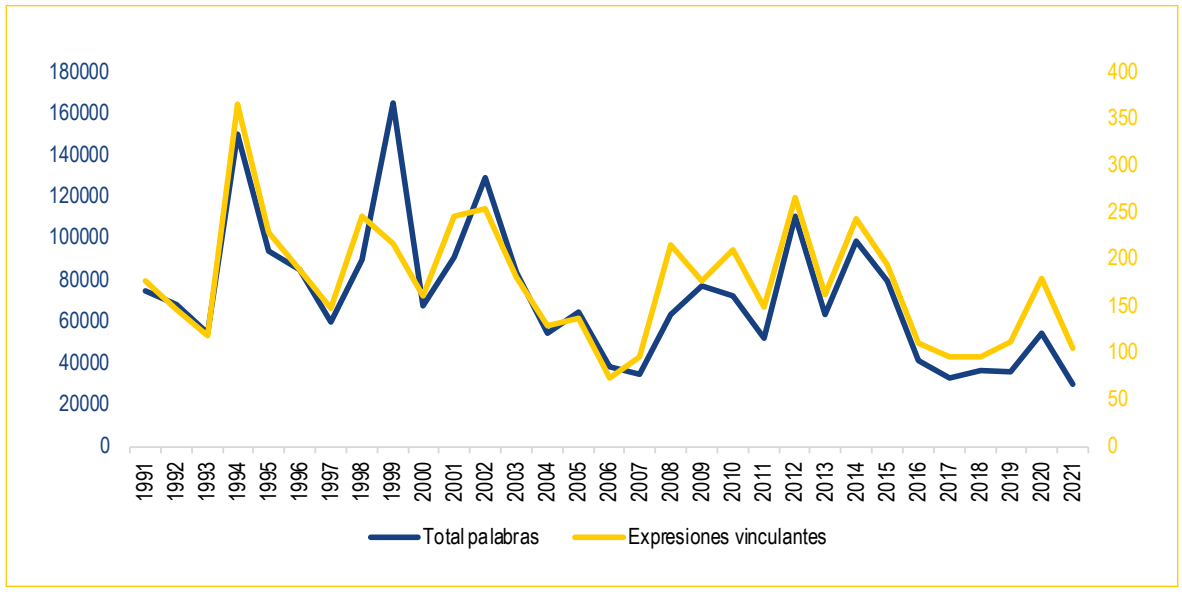
Fuente: elaboración propia con datos OMN-DNP.

Figura 7A -6.
Expresiones vinculantes y total de palabras de las regulaciones, sector de comercio, restaurantes y hoteles



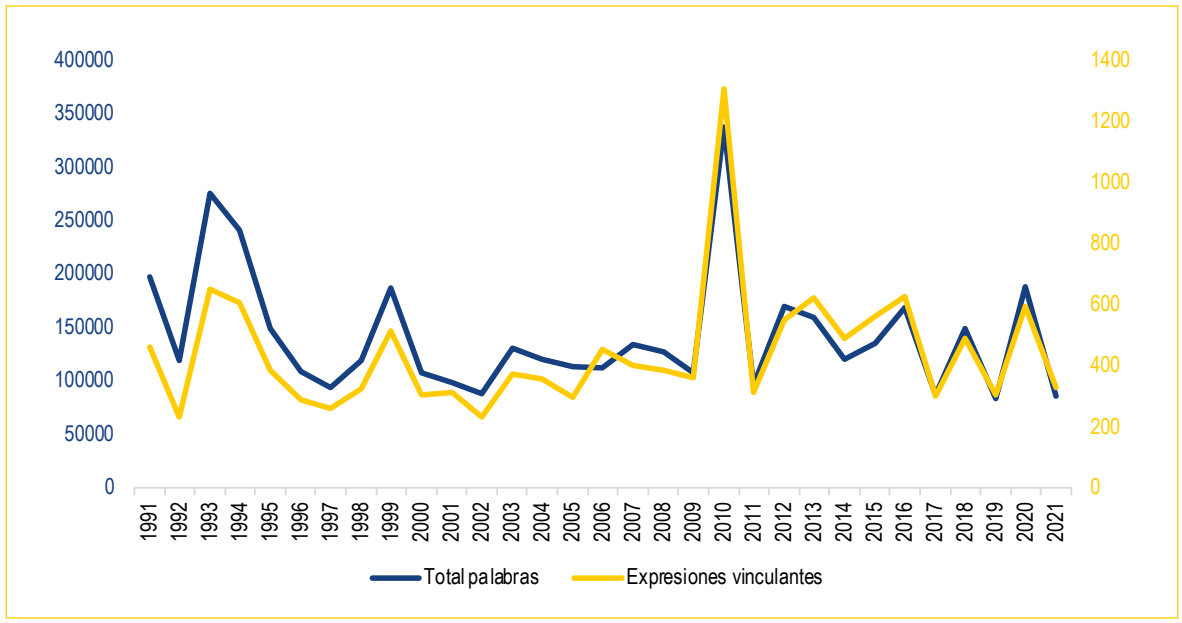
Fuente: elaboración propia con datos OMN-DNP.

Figura 7A -7.
Expresiones vinculantes y total de palabras de las regulaciones, sector de transporte y comunicaciones



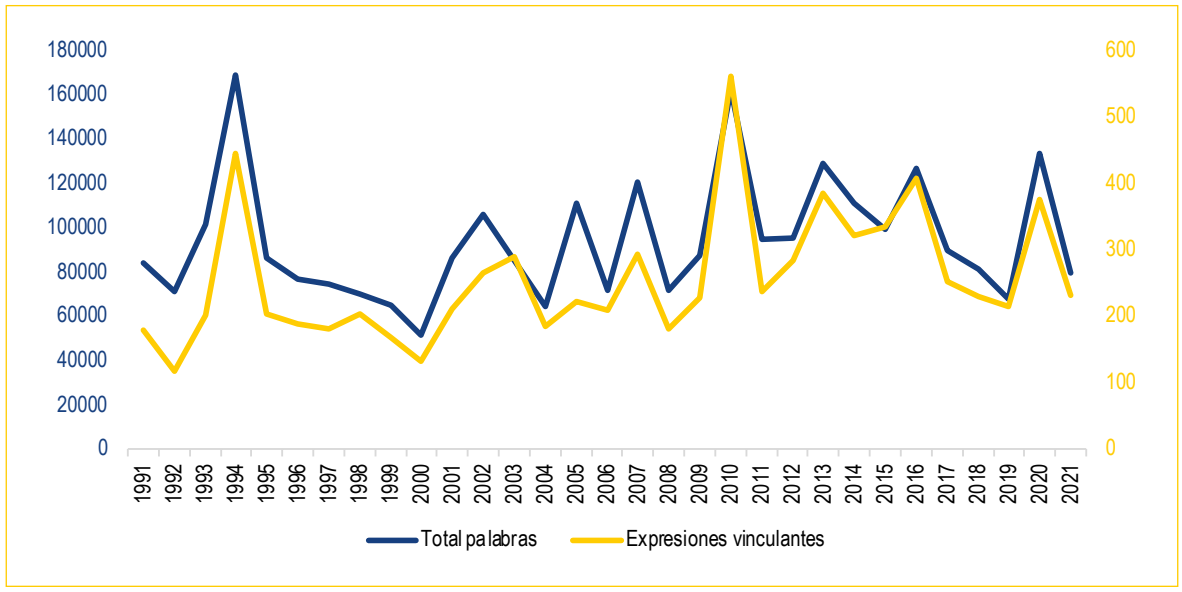
Fuente: elaboración propia con datos OMN-DNP.

Figura 7A -8.
Expresiones vinculantes y total de palabras de las regulaciones, sector de servicios financieros y empresariales



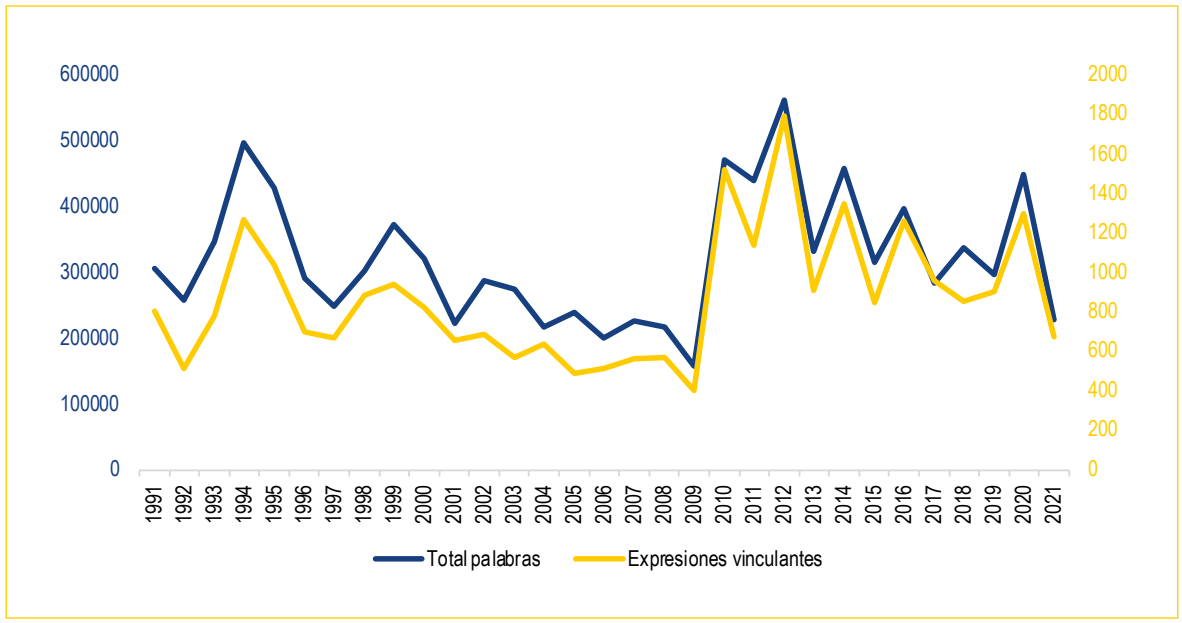
Fuente: elaboración propia con datos OMN-DNP.

Figura 7A -9.
Expresiones vinculantes y total de palabras de las regulaciones, sector de servicios comunales, sociales y personales



Fuente: elaboración propia con datos OMN-DNP.

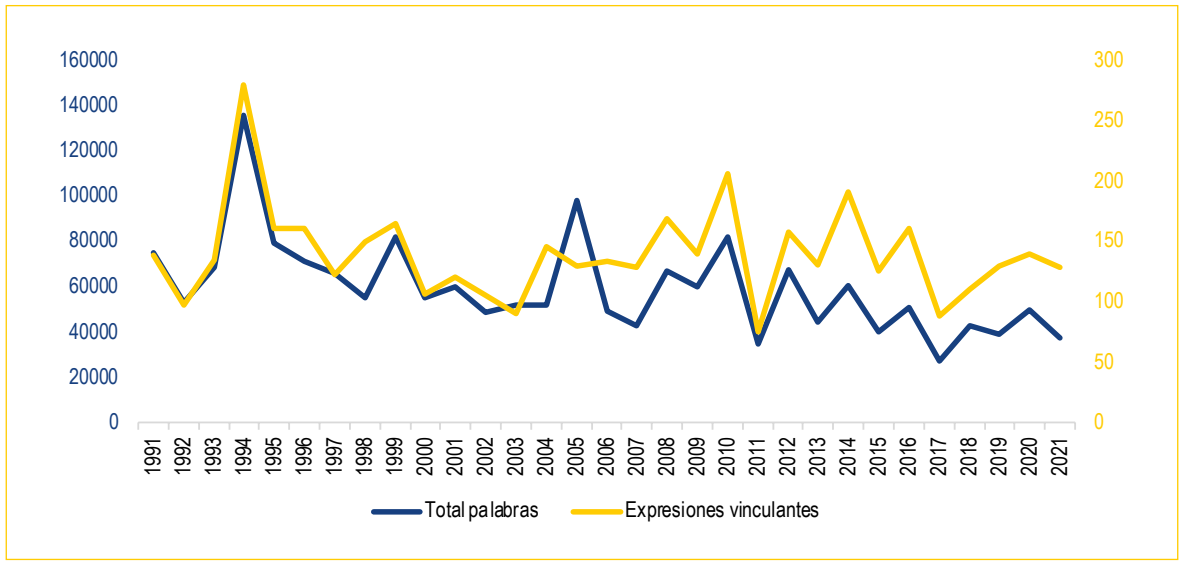
Figura 7A -10.
Expresiones vinculantes y total de palabras de las regulaciones, categoría Otros



Fuente: elaboración propia con datos OMN-DNP.

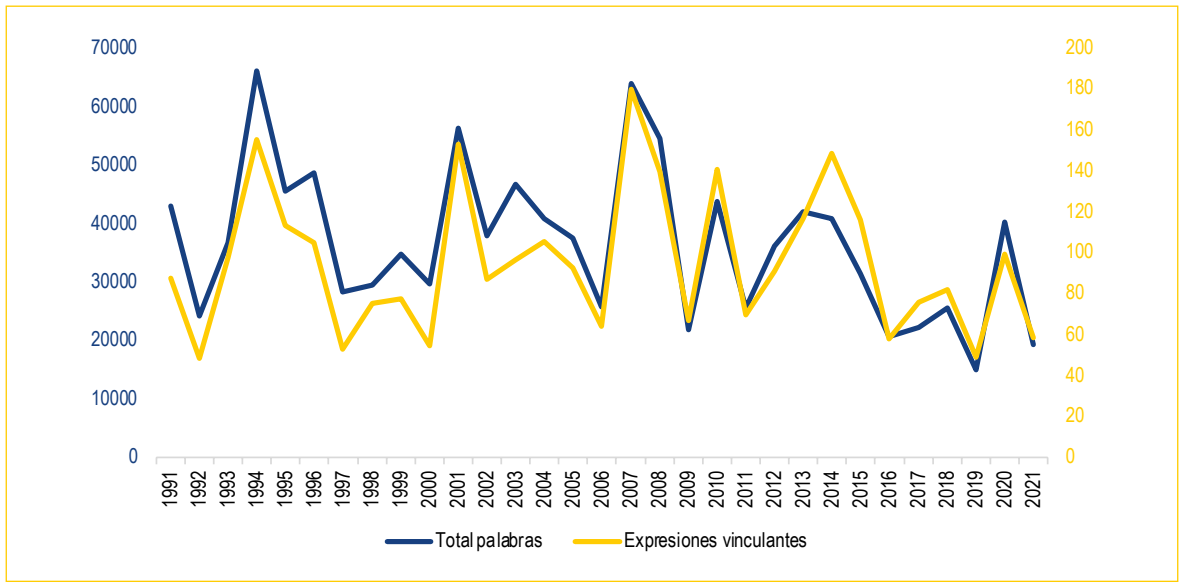
B. Paralelo proporción de expresiones vinculantes ajustadas y no ajustadas por la probabilidad de asociación por sector

Figura 7B -1.
Paralelo proporción de expresiones vinculantes ajustadas y no ajustadas por la probabilidad de asociación al sector agropecuario, silvícola y pesquero



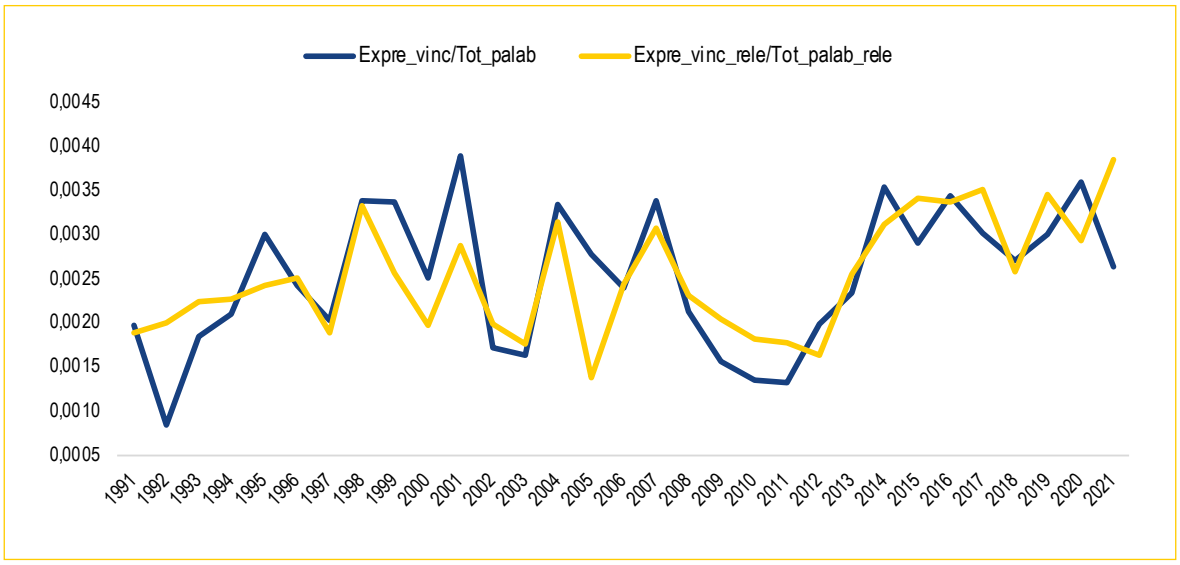
Fuente: elaboración propia con datos OMN-DNP.

Figura 7B -2.
Paralelo proporción de expresiones vinculantes ajustadas y no ajustadas por la probabilidad de asociación al sector de minería



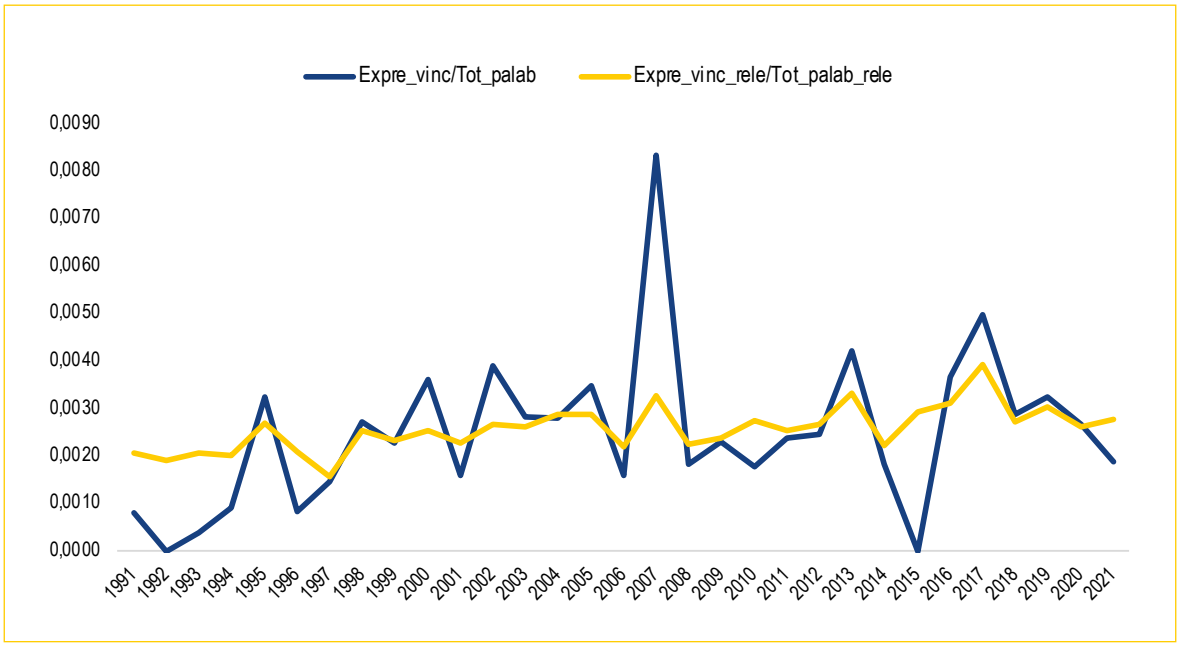
Fuente: elaboración propia con datos OMN-DNP.

Figura 7B -3.
Paralelo proporción de expresiones vinculantes ajustadas y no ajustadas por la probabilidad de asociación al sector de manufactura



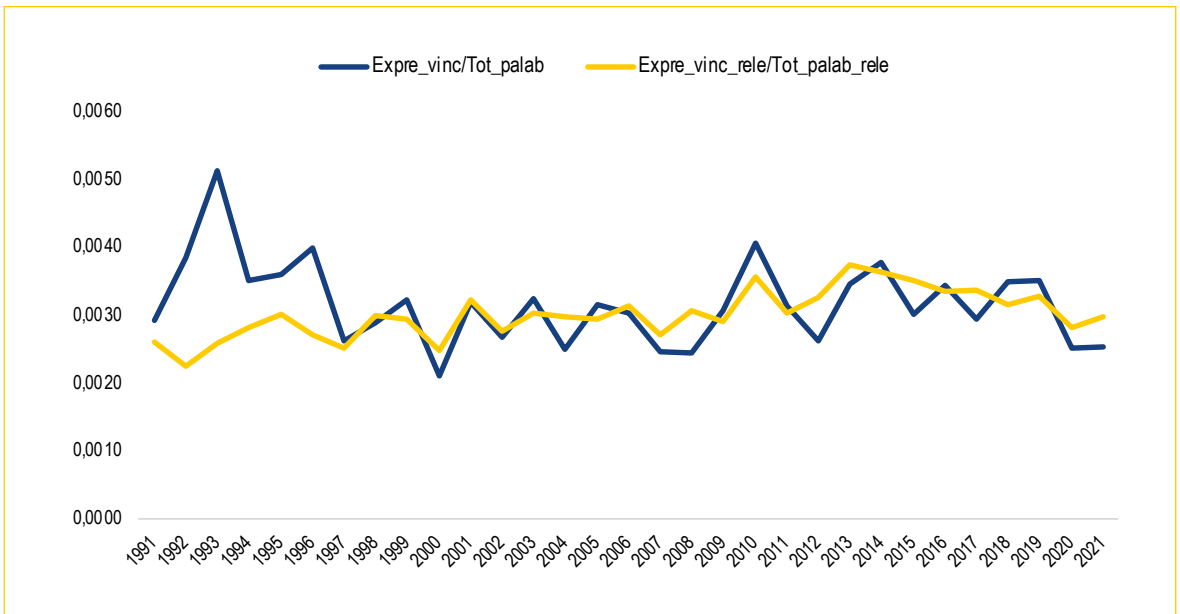
Fuente: elaboración propia con datos OMN-DNP.

Figura 7B -4.
Paralelo proporción de expresiones vinculantes ajustadas y no ajustadas por la probabilidad de asociación al sector de electricidad, gas y agua



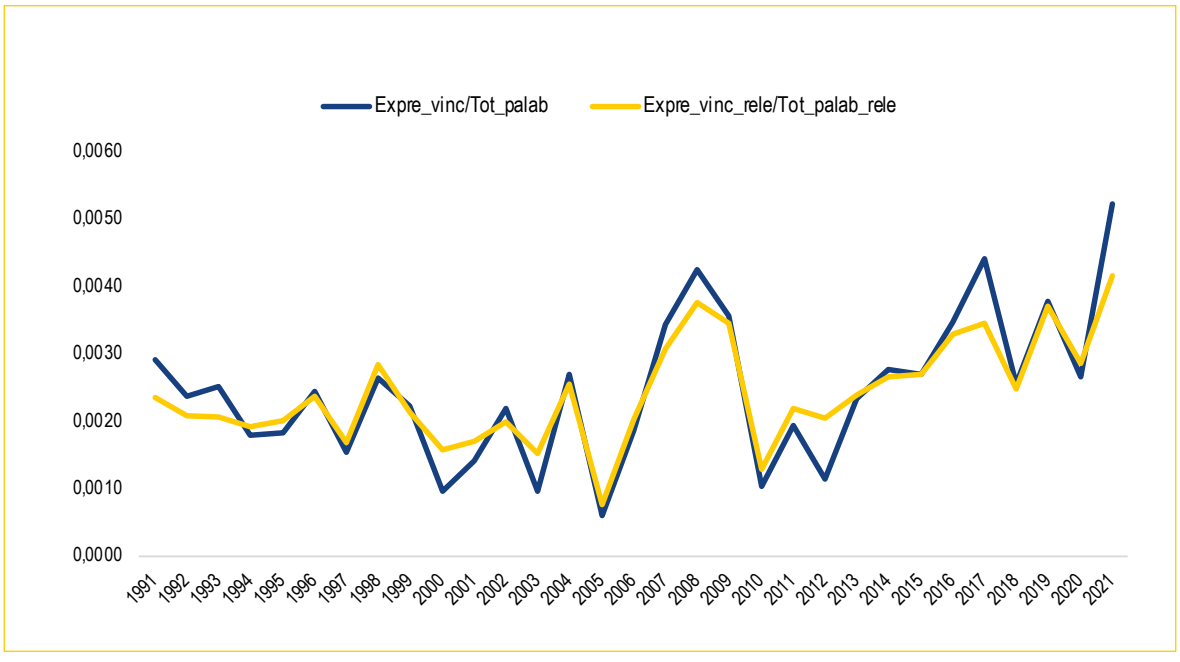
Fuente: elaboración propia con datos OMN-DNP.

Figura 7B -5.
Paralelo proporción de expresiones vinculantes ajustadas y no ajustadas por la probabilidad de asociación al sector de construcción



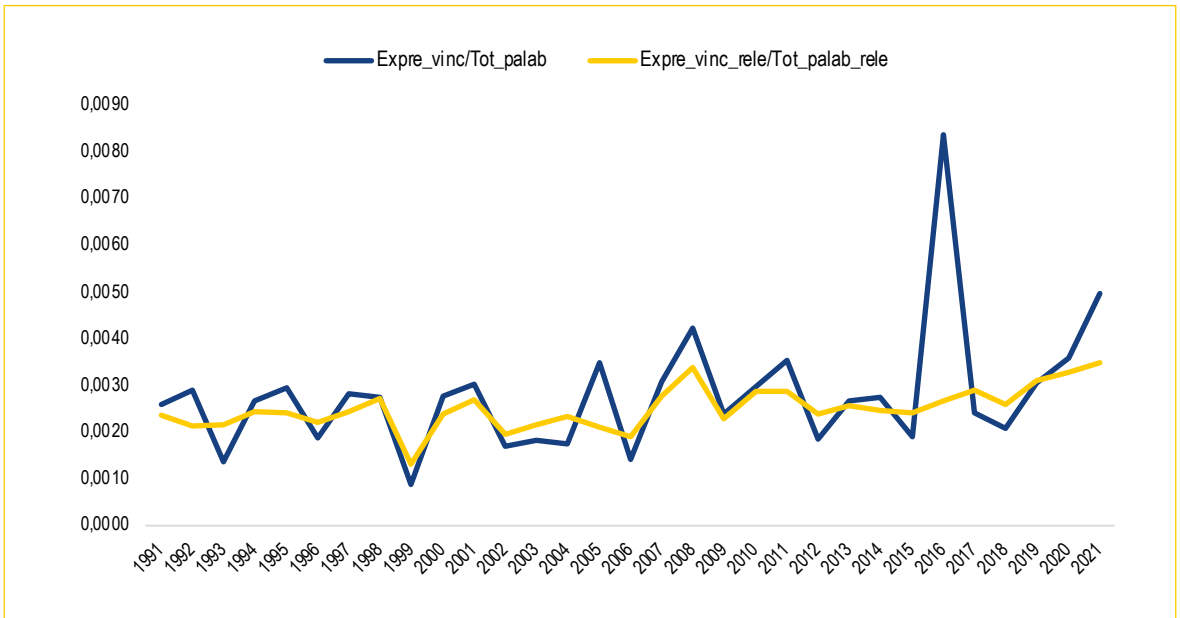
Fuente: elaboración propia con datos OMN-DNP.

Figura 7B -6.
Paralelo proporción de expresiones vinculantes ajustadas y no ajustadas por la probabilidad de asociación al sector de comercio, restaurantes y hoteles



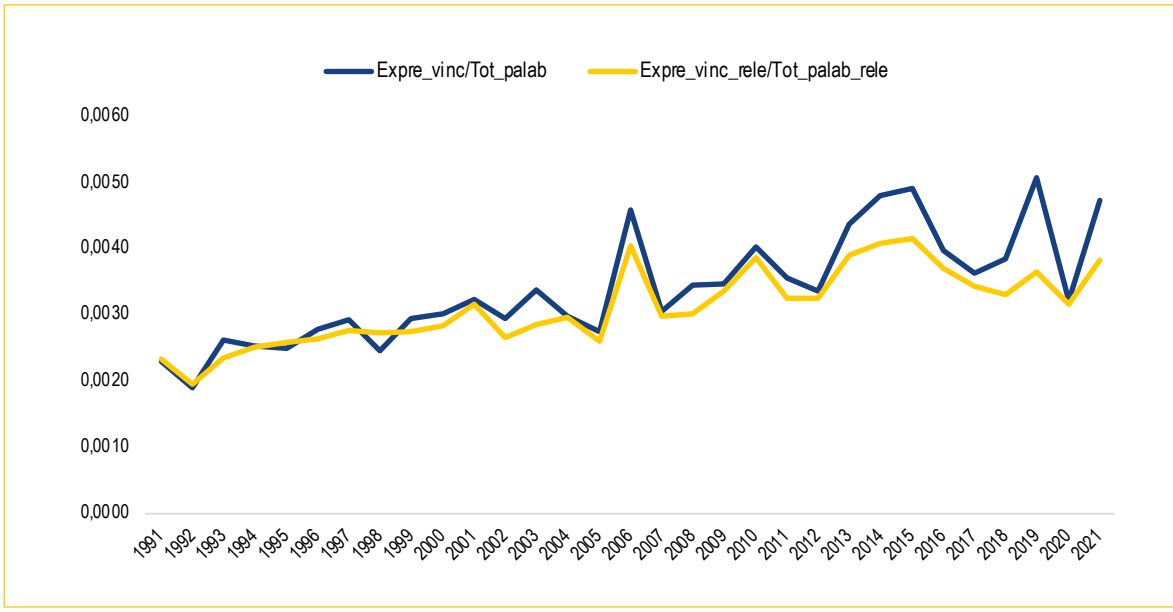
Fuente: elaboración propia con datos OMN-DNP.

Figura 7B -7.
Paralelo proporción de expresiones vinculantes ajustadas y no ajustadas por la probabilidad de asociación al sector de transporte y comunicaciones



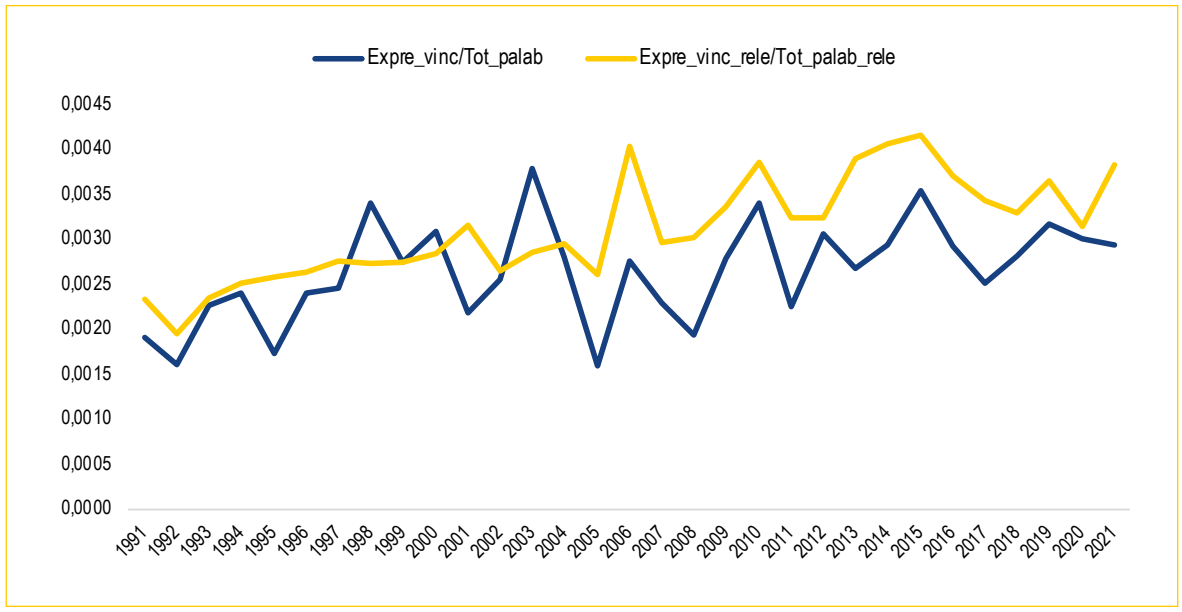
Fuente: elaboración propia con datos OMN-DNP.

Figura 7B -8.
Paralelo proporción de expresiones vinculantes ajustadas y no ajustadas por la probabilidad de asociación al sector de servicios financieros y empresariales



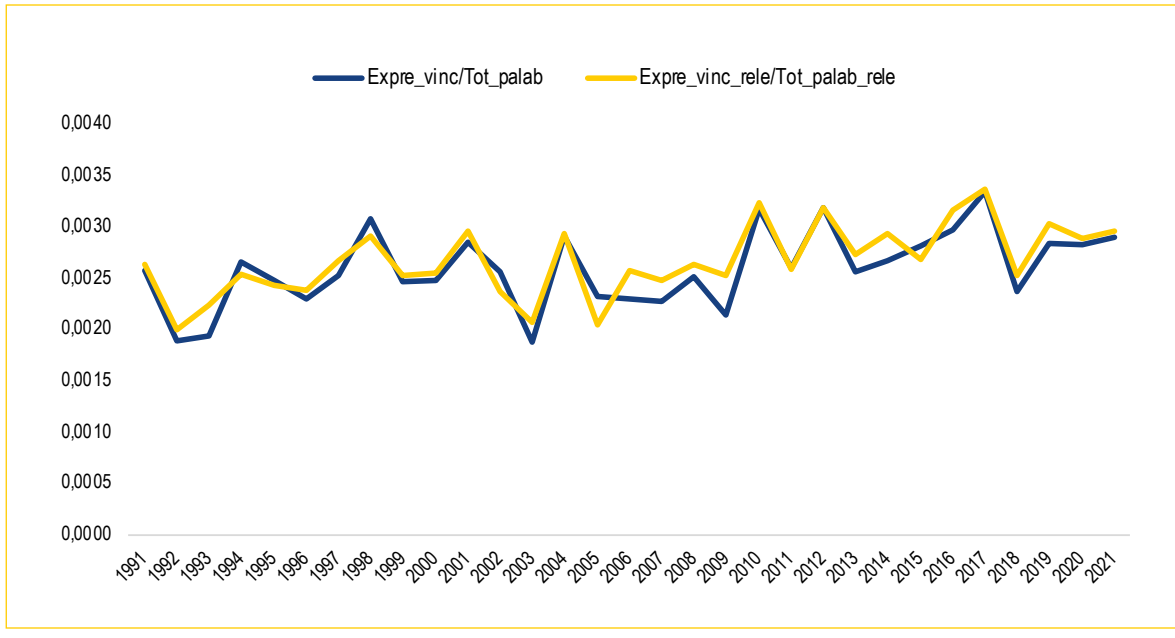
Fuente: elaboración propia con datos OMN-DNP.

Figura 7B -9.
Paralelo proporción de expresiones vinculantes ajustadas y no ajustadas por la probabilidad de asociación al sector de servicios comunales, sociales y personales



Fuente: elaboración propia con datos OMN-DNP.

Figura 7B -10.
Paralelo proporción de expresiones vinculantes ajustadas y no ajustadas por la probabilidad de asociación a la categoría Otros



Fuente: elaboración propia con datos OMN-DNP.

COLOMBIA POTENCIA DE LA VIDA

Departamento Nacional de Planeación - **DNP**

IMPRENTA NACIONAL DE COLOMBIA

Observatorio de Mejora Normativa

Departamento Nacionalde Planeación

Calle 26 núm. 13-19
Edificio ENTerritorio
Bogotá D.C., Colombia
Teléfono: +57 601 3815000

www.dnp.gov.co

Imprenta Nacional de Colombia

Cr 66 núm. 24 – 09
Bogotá, Colombia
Teléfono: +57 601 4578000

www.imprenta.gov.co

DNP Colombia /departamentonacionaldeplaneacion @DNP_Colombia