El empleo es de todos

Mintrabajo

Documentación técnica metodológica

OCUPACOL

Catálogo de Ocupaciones para Colombia

Guía general para la unión de la CUOC con otras fuentes de información de mercado y cálculo de indicadores para OCUPACOL

Contenido

Introducción

El presente documento brinda los lineamientos metodológicos generales necesarios para la obtención de estadísticas relevantes del mercado laboral a nivel de la Clasificación Única Ocupacional de Colombia (CUOC). En particular, la CUOC podría enriquecerse con al menos tres fuentes de información: La Gran Encuesta Integrada de Hogares (oferta laboral/Demanda efectiva), vacantes (SPE, Computrajo, etc. – Demanda insatisfecha) y registros administrativos como el RUES.

Sin embargo, los datos de la GEIH, vacantes y RUES en principio no cuentan con una clasificación de sus observaciones al nivel de CUOC. Por lo tanto, es necesario el desarrollo de una metodología que permita la homogenización y unión de estas diferentes bases de datos con la CUOC.

Este documento describe a grandes rasgos cómo podría unirse la información de las diferentes fuentes de información con la CUOC y los posibles indicadores que se pueden estimar gracias a este cruce. Luego de esta breve introducción, se expone cómo se podría unir la CUOC con la información de la GEIH y cuáles serían los retos implícitos (ver
Figura 1). En la tercera sección se describen los desafíos y la posible forma de cruzar la información de CUOC con vacantes y cómo el RUES se pueden integrar a la CUOC (ver
Figura 2). Finalmente, en la cuarta sección se describen otras fuentes de información y metodologías que podrían utilizarse como insumo a la base COUC enriquecida para la obtención de indicadores novedosos del mercado laboral.

1. Uso de la GEIH para obtener información para la CUOC

La GEIH es la principal fuente de información del mercado laboral. Sin embargo, esta encuesta (hasta el momento) provee información acerca de la ocupación de las personas (ocupados o desocupados) por medio de la Clasificación Nacional de Ocupaciones (CNO) 1970 la cual no posee una correlativa directa con la CIUO - 08 (dado que la estructura de la CUOC sigue la CIUO - 08). Ahora bien, incluso si la GEIH estuviera codificada de acuerdo con la CIUO – 08, no se podría hacer el cruce

de información CUOC-GEIH debido a que (por temas de representatividad) la GEIH estaría disponible, máximo a 4 dígitos de la CIUO – 08, mientras que la CUOC está a 5 dígitos de la CIUO - 08.

Para resolver este desafío, luego de someter el texto de los oficios de la GEIH a un proceso de clasificadores automáticos[1] (ver **Figura 1**, paso 1) propuestos por Cárdenas (2020), se plantea:

- Utilizar una fusión fonética entre la GEIH y la CUOC, dicha metodología utiliza algoritmos computacionales que miden la similitud entre textos y asigna aquel valor que tenga menor distancia, es decir mayor similitud.

Para la fusión fonética de estas bases se utilizaría como insumo a la descripción de la clasificación CUOC (5 dígitos) y por el lado de la GEIH se utiliza el campo de oficio[2] y como un segundo insumo al índice de entrada[3]. Como resultado de este procedimiento se obtendría a la GEIH clasificada con el CUOC (5 dígitos). Asimismo, otra información de la GEIH como el sector, informalidad, salarios, entre otras, se podrían utilizar para refinar este cruce de información. Esta primera fusión tendrá que ser validada por medio del análisis de representatividad de las nuevas ocupaciones asignadas (ver Figura **1**, paso 2).

No obstante, existirán ocupaciones de la CUOC a las cuales este algoritmo no le será posible asignar valores de la GEIH. Por lo tanto, el número de observaciones no clasificadas dependerá de la rigurosidad con la que se implemente la metodología de fusión fonética.

[1] Los clasificadores automáticos utilizan un conjunto de reglas, tales como palabras degradadas, finales de palabras equivalentes, abreviaturas, palabras de reemplazo, alternativas de palabras, etc. para asignarles un código CIUO - 08 a un texto determinado.
[2] Este campo de la GEIH es aquel en donde se registra la respuesta textual del oficio realizado en el puesto de trabajo, es decir, los campos p6370 (para ocupados); y, p7270 y p7330 (para desocupados)
[3] El índice de entrada es uno de los resultados que se obtienen luego de la clasificación automática propuesta por Cardenas (2020) que conceptualmente es un texto alternativo al oficio. Este índice de entrada obtuvo el valor de mayor similitud con alguna ocupación de la CIUO-08 (4 dígitos)

En segundo lugar y posterior a la validación de representatividad de la primera fusión fonética, se pueden tomar a todas aquellas observaciones de la GEIH a las cuales no les fue asignada información de la CUOC para ejecutar la tercera etapa de esta propuesta metodológica (ver **Figura 1**, paso 3):

- Posterior a una revisión manual de los oficios de la GEIH sin clasificar más frecuentes, se podría construir un diccionario de conjuntos de productos y términos coloquiales identificados y que se pueden mapear directamente en los oficios para clasificarlos manualmente.

Con este ejercicio se pueden tener la siguiente serie de indicadores de oferta/demanda efectiva laboral por ocupaciones de la CUOC:

- Ingresos
- Número de empleados
- Tasa de crecimiento del número de empleados
- Top industrias
- Horas trabajadas
- Nivel educativo
- Tasa de informalidad
- Antigüedad – tasa de rotación
- Aglomeración geográfica

1.1. Detalle metodológico

En esta sección se detalla de manera exhaustiva la metodología del pegue de la CUOC con la oferta laboral, es decir con la GEIH (ver **Figura 1**). Grosso modo, esta fase se puede sintetizar en un proceso de pegue a través de la utilización de una fusión fonética de los campos objetivo, que más adelante se detallarán a mayor profundidad, como son las descripciones de las ocupaciones en la CUOC y el oficio textual que provee cada ocupado encuestado en la GEIH.

La metodología de fusión fonética que sea seleccionada o como se mostrará en líneas siguientes, las metodologías fonéticas que sean validadas podrían tomar como insumo campos de la CUOC y GEIH

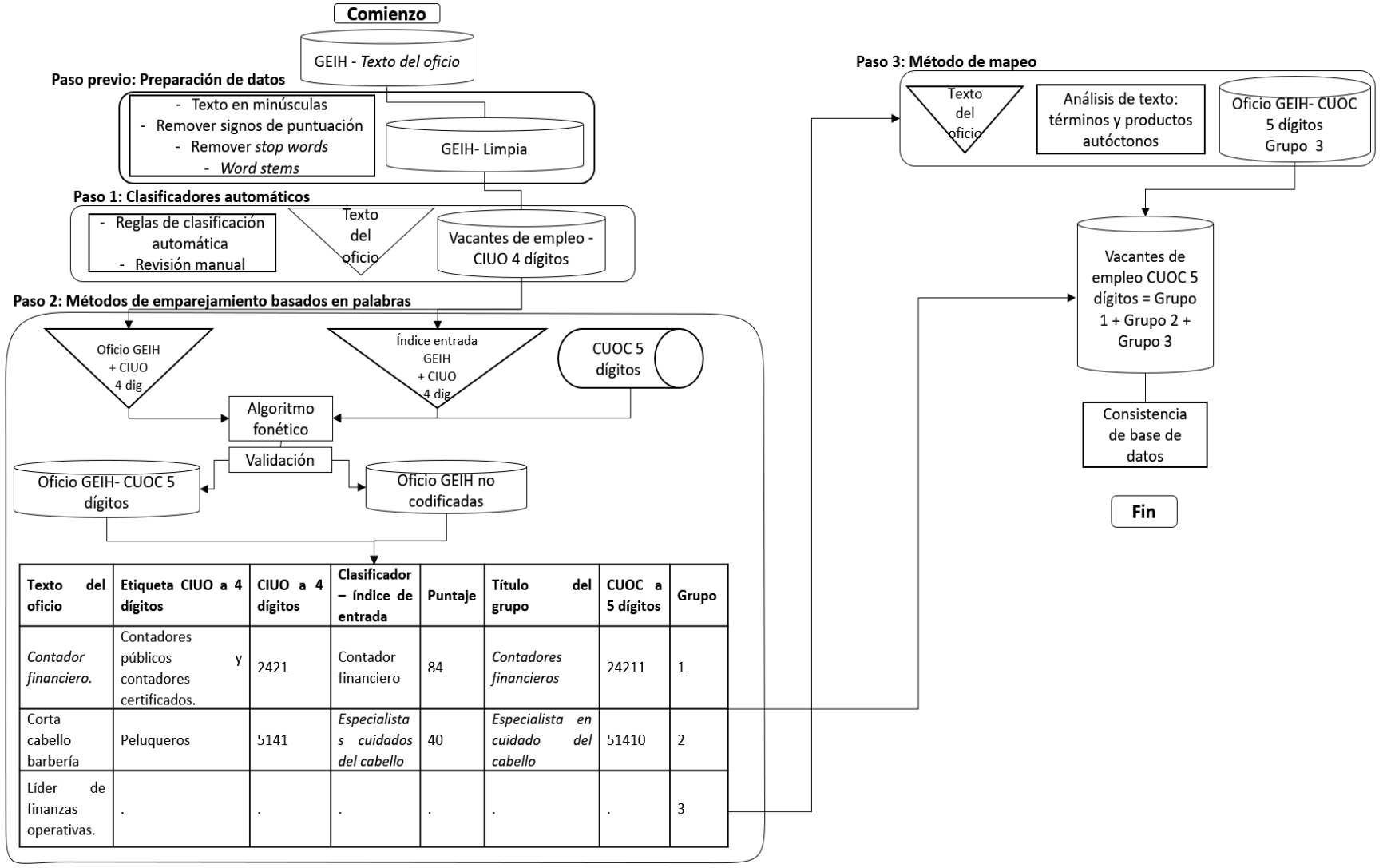
para lograr asignar a cada ocupación CUOC, todas las observaciones de la GEIH que sean más similares con su descripción. Esto con el fin de poder obtener una serie de indicadores de oferta/demanda efectiva laboral para cada ocupación de la CUOC (5 dígitos).

1.1.1. La fuente de información - Gran Encuesta Integrada de Hogares-GEIH

Como se puede observar en la

Figura 1, para llevar a cabo la primera fase se podrían utilizar dos fuentes de información. Por un lado, se podría utilizar la base CUOC que brinda el detalle de 670 ocupaciones en su nivel más desagregado (5 dígitos). Por otro lado, en Colombia, la fuente de información que recoge el pulso de la demanda laboral efectiva es la GEIH que se utilizaría para completar los insumos de la metodología de la fusión fonética. Así, se podría obtener como resultado una base CUOC asignada un conjunto de observaciones de la GEIH, las cuales serían las más semejantes bajo los criterios de la fusión fonética.

Figura 1. Esquema general para codificar los datos de GEIH a nivel de 5 dígitos de la CUOC



Texto del oficio	Etiqueta CIUO a 4 dígitos	CIUO a 4 dígitos	Clasificador – índice de entrada	Puntaje	Título del grupo	CUOC a 5 dígitos	Grupo
Contador financiero.	Contadores públicos y contadores certificados.	2421	Contador financiero	84	*Contadores financieros*	24211	1
Corta cabello barbería	Peluqueros	5141	*Especialistas cuidados del cabello*	40	*Especialista en cuidado del cabello*	51410	2
Líder de finanzas operativas.	.	.	.	.	.	.	3

La GEIH es una encuesta con periodicidad mensual que tiene como objetivo principal proporcionar información esencial sobre el tamaño y estructura de la fuerza de trabajo del país debido a que se aplica en todo el territorio nacional y que permite la desagregación de resultados para el total nacional, total cabeceras, total centros poblados y rural disperso, cada una de las 23 ciudades capitales y áreas metropolitanas, y San Andrés. Además, el DANE menciona que:

> *La Gran encuesta integrada de hogares es una encuesta mediante la cual se solicita información sobre las condiciones de empleo de las personas (si trabajan, en qué trabajan, cuánto ganan, si tienen seguridad social en salud o si están buscando empleo), además de las características generales de la población como sexo, edad, estado civil y nivel educativo, se pregunta sobre sus fuentes de ingresos.*

Actualmente, esta encuesta provee información acerca de la ocupación de las personas (ocupados o desocupados) por medio de la Clasificación Nacional de Ocupaciones (CNO) 1970 la cual no posee una correlativa directa con la CUOC, este hecho sucede principalmente ya que la CUOC sigue la estructura de la CIUO – 08. No obstante, incluso si la GEIH estuviera codificada bajo el paraguas de la CIUO – 08, no se podría hacer el cruce de información CUOC-GEIH directamente debido a que la GEIH estaría disponible a lo máximo a 4 dígitos de la CIUO – 08, mientras que la CUOC está a 5 dígitos de la CIUO - 08.

El campo de oficio de la GEIH podría ser aquella pieza clave para llevar a cabo el pegue fonético, en este campo se registra la respuesta textual de los encuestados, referente al oficio que realizan o realizaron, para el caso de los desocupados, estos son los campos p6370 (para ocupados); y, p7270 y p7330 (para desocupados) de la GEIH. Como resultado de este procedimiento se obtendría a la GEIH clasificada con el CUOC (5 dígitos).

Por lo expuesto, para el mapeo de las ocupaciones a las que pueden pertenecer cada uno de los ocupados y desocupados de la GEIH, la utilización del campo del detalle del oficio que realizan sería el más pertinente. No obstante, es preciso denotar algunas de las limitantes que podrían implicar su utilización:

- Una particularidad de este campo se basa en que es una muy breve descripción que proporciona el encuestado al momento de responder su encuesta que, a diferencia, por ejemplo, de los campos título o descripción en la base de vacantes, tendrían textos más cortos. Este hecho podría desempeñar un papel clave en la implementación de la metodología del pegue fonético ya que es más probable que desempeñe una mejor clasificación conforme la metodología tenga más información, pero no cualquier clase de información adicional sino aquella que sea significativa para alcanzar un mayor número de registros clasificados correctamente.

  A modo de ejemplo se puede mostrar que en algunas ocasiones es sumamente necesario más información, tal es el caso de una respuesta como “profesor”, para esta respuesta el algoritmo fonético no podrá realizar una clasificación óptima, basado en el hecho que existen diferentes descripciones CUOC que tienen incluido este término por ejemplo “Profesores de instituciones de educación superior”, “Profesores de educación secundaria”, entre otros. Entonces, con la información provista no le sería posible al algoritmo decidir entre cuál de estas opciones sería correcta la asignación.
- El uso de términos coloquiales como “vendedor de empanadas” puede ser causa de una no clasificación. En este mismo sentido, para el algoritmo podría representar ambiguo el uso indistinto entre términos que en el lenguaje común son usados indistintamente para el mismo fin, por ejemplo: docente de educación superior , educador de educación superior, maestro de educación superior y profesor de educación superior que, pese a que están siendo acompañadas de mayor información como el nivel al cual dicta clases (superior), la CUOC es estricta en señalar que la descripción correspondiente es “Profesores de instituciones de educación superior”

Por lo expuesto, otra información de la GEIH como el sector, informalidad, salarios, entre otras, podrían ser utilizados como criterios complementarios a este cruce. No obstante, esta primera fusión tendrá

que ser validada por medio de un análisis de consistencia de la base como un análisis de robustez de la clasificación.

1.1.2. Método de Fusiones fonéticas

Entre los algoritmos que se han desarrollado dirigidos hacia el análisis de texto se encuentra la familia de fusiones fonéticas que van desde los algoritmos más simples de búsquedas exactas, es decir, un patrón dentro de una palabra hasta analizar en qué proporción o cuánto dista una palabra a ser igual a otra. Partiendo desde lo más general se encuentra la búsqueda de una ocurrencia exacta (coincidencia) o encontrar una coincidencia no exacta utilizando caracteres con un significado especial, por ejemplo, mediante expresiones regulares.

También se han desarrollado algoritmos, algo más complejos, en función de cómo suenan las palabras, en otras palabras, son iguales, pero están escritas de una manera diferente. Un ejemplo común es la búsqueda de parientes utilizando los apellidos de un registro que muchas veces pueden incurrir en faltas ortográficas pero que, de cierto modo, suenan de manera similar

Por último, pero no menos importante, se encuentran los algoritmos especializados que identifican cuántos cambios o ediciones son necesarios para pasar de una palabra a otra. Conforme sean menores estas ediciones, mayor será el nivel de similitud. En suma, existen 3 grandes grupos de algoritmos que realizan fusiones fonéticas: búsquedas exactas, algoritmos fonéticos y fusiones por cambios o ediciones. Dentro de estos grupos, los algoritmos más populares son: cadenas de Boyer-Moore, expresiones regulares, n-gramas, Soundex y distancias de Levenshtein, Hamming y Jaro-Winkler. A continuación, se detallarán los principales algoritmos fonéticos y sus características:

Expresiones regulares

Las expresiones regulares son una secuencia de caracteres que se pueden usar para definir un patrón de búsqueda para encontrar texto. Para usar expresiones regulares se especifica las reglas para el conjunto de posibles cadenas que desea hacer coincidir y luego se hace preguntas como ¿Esta cadena coincide con el patrón? o ¿Hay alguna coincidencia con el patrón en cualquier lugar de esta

cadena? ". Finalmente, también puede utilizar expresiones regulares para modificar una cadena o dividirla de varias formas.

N-gramas

En el procesamiento del lenguaje natural se conoce a los N-gramas como cadenas de palabras, donde n representa la cantidad de palabras que está buscando. Este algoritmo encontró su aplicación principal en un área de modelos lingüísticos probabilísticos. Debido a que estiman la probabilidad de ocurrencia del siguiente elemento en una secuencia de palabras.

En la modelización de lenguaje asume una estrecha relación entre la posición de cada elemento en una cadena, calculando la ocurrencia de la siguiente palabra con respecto a la anterior. En particular, el modelo de N-gramas determina la probabilidad de la siguiente manera - N-1. Por ejemplo, un modelo de trigrama (con N = 3) predecirá la siguiente palabra en una cadena basándose en las dos palabras anteriores como N-1 = 2.

Distancia de Levenshtein

El valor de la distancia describe el número mínimo de eliminaciones, inserciones o sustituciones que se requieren para transformar una cadena en otra. Cuanto mayor es la distancia de Levenshtein, mayor es la diferencia entre las cadenas de texto. Por ejemplo, de "gato" a "gato", la distancia de Levenshtein es 0 porque tanto la cadena de origen como la de destino son idénticas, no se necesitan transformaciones. Por el contrario, de "gato" a "gala", la distancia de Levenshtein es 2: se deben hacer dos sustituciones para convertir "gato" en "gala".

Limitaciones

Como se ha descrito en la subsección anterior, el uso de fusiones fonéticas sería pertinente dadas las características de las bases de CUOC y GEIH para implementar la primera fase de clasificación. En este punto es preciso señalar las potenciales limitaciones del método y que darían origen a una validación entre los métodos que se explicitará en la siguiente subsección que podría ser acompañada de una revisión manual para identificar de una manera más minuciosa, por ejemplo:

- Sinónimos

Las diferentes fusiones fonéticas tienen una capacidad limitada a realizar clasificaciones únicamente con base en los inputs que se les proporcione sin embargo no consisten en métodos de aprendizaje que potencialmente tendrían la capacidad de identificar sinónimos tales como: los términos “Asistente” y “Auxiliar” se consideran categorías diferentes, aunque pueden, en muchas ocasiones, referirse a la misma categoría de trabajo.

- Idiomas

Es probable que dentro de las descripciones de los oficios que contenga la GEIH los haya respondido una persona no hispanohablante o una persona colocó la posición ocupacional que tiene dentro de la empresa donde trabaje y estos puedan estar en otro idioma. Este particular también se sale de los límites que abarcan las fusiones fonéticas.

1.1.3. Clasificadores automáticos y fusiones fonéticas en la práctica

Para la real implementación de los clasificadores automáticos y algoritmos de fusión fonética que se detallaron anteriormente es indispensable un preprocesamiento de las fuentes de información, especialmente de los campos objetivo que se utilizarán de cada de estas como input del algoritmo fonético (el campo descripción de la ocupación en la CUOC y el campo oficio textual en la GEIH).
Este preprocesamiento se ha establecido como norma inicial para la implementación cualquier algoritmo que tenga implícito un análisis de texto y va enfocado principalmente a homologar los textos a través de la eliminación de términos, signos y demás palabras “basura” que al ser incluidas solo generan ruido para la óptima clasificación (ver
Figura 1, paso previo). Así, entre los pasos iniciales se tiene:

- Conversión a minúsculas, dependiendo de la sensitividad con la que vayan a ser implementados los algoritmos en algunos de estos podría resultar que, por ejemplo, hablando en sentido estricto, el término “jefe” podría ser distinto a “JEFE”, lo cual sería incorrecto.
- Eliminación de signos de puntuación, un ejemplo, en sentido estricto podría decir que “ingeniero, computación” es diferente a “ingeniero computacion”, esto sería incorrecto.
- Eliminación de stopwords, palabras sin significado como artículos, pronombres, preposiciones, etc. que incluyen ruido en los algoritmos de clasificación y que, en el caso de

no implementarse, podría resultar que “ayudante *de* cocina” sea diferente a “ayudante *en la* cocina”.

- Raíces de términos, este es un método que reduce cada palabra a su raíz terminológica. Este método es indispensable ya que convertiría tanto a “ayudante bibliotecario” como a “ayudante biblioteca” en “ayudante bibliotec” que resultaría en una clasificación unificada y acertada.

Luego de este preprocesamiento ingresarían los inputs del algoritmo fonético limpio y sin ruidos para que se pueda dar lugar la implementación de una clasificación más pura.

1.1.4. Algoritmo de validación

Por lo expuesto, cada uno de los tipos de fusiones fonéticas que existen tienen ventajas que conjuntamente podrían ser una potente herramienta que clasifique a gran parte de las observaciones de la GEIH, dejando por fuera a una cantidad minúscula que debería pasar a la tercera fase que se propone en este documento (ver
Figura 1, paso 3). En función de lo expuesto, este documento propone la formulación de un algoritmo que funcione como un *trade off* y mida el beneficio entre optar por fusiones fonéticas cuyos índices de clasificación sean altos por aquellas que se obtengan índices bajos de clasificación.

Este algoritmo podría estar sustentado en un símil del concepto de costo-oportunidad ya que debería tener en cuenta en su construcción las ventajas que se estrían ganando al optar por ciertos tipos de fusiones por abandonar las características que nos pudieran ofrecer los algoritmos que obtuvieron índices de clasificación bajos.

1.1.5. Método de mapeo: títulos que no clasificaron en la fusión fonética

Luego de identificar a los oficios de la GEIH que no fueron clasificados en las 2 primeras etapas de la propuesta metodológica se procede a la construcción de una lista de términos a través de un rastreo

manual de aquellos productos[4], acciones[5], lugares de trabajo[6] que son característicos del mercado laboral local, pero, debido a sus nombres coloquiales o particularidades en la jerga regional, no se lograron emparejar al inicio de la metodología.

Tras haber estructurado estas listas de términos se realiza un proceso de mapeo de las posibles combinaciones de estos, a la vez se asigna un código CUOC manualmente a los oficios GEIH donde sea afirmativa la localización de estos términos.

Una sustancial proporción de esta submuestra se clasifica por medio de un proceso diferenciado que además utiliza la variable p6880 de la GEIH. Este proceso tiene como finalidad categorizar a los vendedores ambulantes de comida. De este modo, si en el texto del oficio se logra mapear al término venta o sinónimos, algún tipo de comida y su lugar de trabajo es un sitio al descubierto en la calle se le asigna el código CUOC 52120 que corresponde a Vendedores ambulantes de alimentos preparados para consumo inmediato.

1.1.6. Validación de la consistencia de la base

Para comprobar la robustez de la clasificación y la funcionalidad del algoritmo fonético es preciso realizar ejercicios que identifiquen cuáles grupos ocupacionales CUOC podrían ser "representativos" en función del número de observaciones de GEIH que se les fue asignados.

La CUOC está compuesta de 449 grupos primarios, 136 subgrupos, 43 subgrupos principales y 10 grandes grupos, según la similitud en cuanto al nivel de competencias requeridas para los empleos y la especialización de estas. Esto permite la producción de estadísticas detalladas, resumidas y comparables a nivel internacional. Para continuar manteniendo esta comparabilidad con los datos que

[4] Algunos productos alimenticios de preparación y venta son: empanadas, tamales, arepas, aromáticas, café tinto, mazamorra, peto, lechona, pincho, jalea, oblea, bollo, envuelto, fritanga, etc.
Algunos productos de cultivo son: caña, papa, plátano, yuca, maíz, fríjol, mora, etc.
[5] Principalmente acciones de servicio doméstico como lavar y planchar ropa. Actividades de apoyo a la construcción como estucar paredes; pegar ladrillo o baldosa; revolver y hacer mezclas. Actividades de peluquería como arreglar uñas, cortar cabello, etc.
[6] Lugares de trabajo como café internet o coductores de vehículos, colectivos, furgones, etc.

nos pueda proporcionar el pegue con la GEIH, en este documento se propone que, a partir del número de observaciones que se hayan asignado a cada grupo ocupacional COUC 5 – dígitos hallar el coeficiente de variación entre estos, dentro de cada grupo CUOC 4 – dígitos al que pertenezcan. Luego de la obtención de este coeficiente de variación tal como sugiere Cárdenas (2020) se propone crear intervalos alrededor de este con el 50% de la mediana. Finalmente, aquellos grupos CUOC que queden por fuera de estos intervalos se los podría considerar como una muestra no consistente para la obtención de estadísticas.

1.1.7. Posibles ejercicios futuros

El uso de modelos de máquinas que clasifican los títulos de trabajo en códigos de ocupación ha surgido en las últimas décadas. Gweon et al. (2017) destacan que instituciones estadísticas de Australia han favorecido este método. En términos concretos, el aprendizaje automático es un "conjunto de métodos que pueden detectar automáticamente patrones en los datos y luego usar los patrones descubiertos para predecir datos futuros o para realizar otros tipos de toma de decisiones en condiciones de incertidumbre" (Murphy 2012, p. 1) . Además, como señala Murphy (2012), la clasificación (aprendizaje supervisado) es, probablemente, la forma de aprendizaje automático más utilizada para resolver problemas del mundo real.

La idea de este método es clasificar un "documento", para este caso partículas podrías ser un puesto de trabajo, en una de varias clases (C) basándose en algunos insumos de capacitación aprendidos previamente (X). La computadora determina cómo clasificar un documento basándose tanto en un conjunto de datos de entrenamiento como en un algoritmo de asociación particular. El primero se refiere a un conjunto de datos preprocesados con un número N de ejemplos de entrenamiento (D).
Por lo tanto, luego de la obtención de la validación de base CUOC, esta podría ser un potente insumo que funcione, como se dijo en líneas anteriores, como una base de entrenamiento de modelos de aprendizaje automático y de esta manera tratar de cerrar esa brecha de limitaciones que conllevan la utilización única de fusiones fonéticas.

2. Uso de la información del servicio público de empleo de vacantes para la CUOC

Una clasificación es un marco teórico utilizado para agrupar trabajos en función de las tareas y deberes realizados en el trabajo. Este marco teórico es un elemento clave para realizar análisis estadísticos estandarizados sobre el mercado laboral. Lo anterior, debido a que permite medir, por ejemplo, los tipos de trabajos (complejidad de tareas) demandados u ofertados en un país. Para mantener la clasificación ocupacional, el DANE en Colombia utiliza la Clasificación Nacional de Ocupaciones CNO, sin embargo, al no tener una comparabilidad internacional en Cárdenas (2020) se propone la utilización de la CIUO – 08.

El nivel actual de desagregación de la CIUO – 08 a 4 dígitos no proporciona suficientes detalles para algunos usuarios (p. ej. Los programas vocacionales y de capacitación necesitan un mayor nivel de desagregación para proporcionar adecuadamente las habilidades requeridas para un determinado trabajo).

En consecuencia, el DNP y el Ministerio de Trabajo ha comenzado a desarrollar una versión más detallada de la CIUO - 08, denominada “Clasificación Única de Ocupaciones para Colombia, CUOC”. Este proyecto tiene como objetivo, desagregar la CIUO - 08 de 4 dígitos en un quinto nivel. El quinto nivel permite identificar con más detalle los grupos ocupacionales que comparten tareas similares. La CUOC está compuesta de 449 grupos primarios, 136 subgrupos, 43 subgrupos principales y 10 grandes grupos, según la similitud en cuanto al nivel de competencias requeridas para los empleos y la especialización de estas. Esto permite la producción de estadísticas detalladas, resumidas y comparables a nivel internacional.

Entre las diferentes bases de datos que pueden clasificarse potencialmente como CUOC de 5 dígitos, la información de los portales de empleo se destaca como una fuente prometedora, donde la extensión de la CUOC podría ser utilizada para revelar información sobre el mercado laboral de Colombia. La información del portal de empleo ha atraído notablemente la atención de investigadores y hacedores de política pública, debido a que es una fuente que puede ofrecer (a bajo costo) información detallada

sobre la demanda laboral, que de otro modo sería complejo recopilar por otros medios. Esta información se puede utilizar, por ejemplo, para informar a los proveedores de formación y educación, sobre las habilidades necesarias para un conjunto de trabajos. Es importante destacar que, dado el nivel de detalle de variables como los títulos y las descripciones de los puestos, los datos de los portales de empleo pueden codificarse a un nivel de 5 dígitos de la CUOC.

Sin embargo, como lo mencionaron Cardenas-Rubio (2020) y Kureková et al. (2014), los datos de los portales de empleo no están codificados y requieren un esfuerzo considerable para organizarlos en aras de realizar un análisis estadístico. Diferentes empresas como Burning Glass, Emsi, etc, han comenzado a recopilar y organizar información sobre vacantes de fuentes en línea. Estos datos han sido valiosos para identificar habilidades y demanda de ocupaciones en Colombia. A pesar de lo anterior, algunas de estas compañías han codificado la información de vacantes con el CIUO a un nivel de 4 dígitos (en el mejor de los casos), mientras que otras utilizan diferentes clasificaciones y agregaciones. Además, en la mayoría de los casos, las empresas no revelan cómo se recopiló y organizó la información, lo que dificulta la evaluación del alcance y la coherencia de los datos de vacantes en diferentes niveles de desagregación (por ejemplo, en los niveles de 5 dígitos de la CUOC).

Para efectos del análisis propuesto, las bases de datos de portales web de empleo son las propuestas por Cárdenas (2020) que recolectan información de Computrabajo y Elempleo desde 2018 hasta 2021. Estos portales son significativos en términos del número de visitas por día, y del volumen de información disponible. Esta base de datos es codificada para la CIUO-08 de Colombia a 4 dígitos y la CUOC a 5 dígitos, utilizando principalmente los clasificadores automáticos propuestos por Cardenas (2020). La cual es, una herramienta para clasificar en grupos ocupacionales los títulos de trabajo (Jones & Elias, 2004). Además, se lleva a cabo técnicas de text-mining (como fuzz merge, patrones de identificación, entre otras) para estandarizar variables como habilidades y requisitos de calificación, salarios, etc., que complementan los datos al nivel de 5 dígitos de la CUOC.

Por lo tanto, los datos recopilados y el uso de herramientas de clasificación son un buen punto de partida para evaluar la posibilidad de codificar la información de los portales de trabajo a nivel de 5 dígitos de la CUOC. El propósito de esta sección es explicar cómo se podría codificar la base de datos

de vacantes a nivel de 5 dígitos de la CUOC y proponer un enfoque estadístico para evaluar la coherencia de los datos de vacantes a nivel de la CUOC a 5 dígitos. Así, el presente documento está dividido en siete subsecciones. La primera sección es la presente introducción; la siguiente sección, presenta una descripción general de metodología propuesta; la tercera sección, explica cómo se pueden utilizar los resultados de las herramientas de clasificación utilizadas en Cardenas (2020) para codificar la información de los portales de empleo a nivel de 5 dígitos; la cuarta sección utiliza métodos de emparejamiento basados en palabras que utilizan títulos de trabajo y resultados de las herramientas de clasificación para codificar las observaciones de vacantes en la CUOC 2020 a nivel de 5 dígitos.

Debido a que es posible que existan títulos de trabajo "difíciles de codificar"; la quinta sección, propone mapear la información de habilidades en la base de datos de vacantes; la sexta sección, propone un algoritmo de *machine learning* que utiliza los títulos de los trabajos y las habilidades para codificar las observaciones de vacantes en el CUOC a nivel de 5 dígitos; la última sección sugiere un método para probar la consistencia de la base de datos de vacantes para la inferencia estadística de la CUOC a nivel de 5 dígitos.

2.1. Resumen general para codificar la información sobre las vacantes en la CUOC a nivel de 5 dígitos.

Como se mencionó anteriormente, la información de los portales de empleo contiene diferentes variables que podrían permitir la codificación ocupacional a nivel de 5 dígitos. Particularmente, la variable título de trabajo es la columna vertebral para codificar la base de datos de vacantes al CUOC a 5 dígitos. Esta variable resume las tareas requeridas para un determinado puesto de trabajo, por lo tanto, sirve como primer y principal ingrediente para conocer el código ocupacional de una determinada observación en la base de datos de vacantes (Cárdenas-Rubio, 2020). El segundo ingrediente, corresponde a las habilidades demandadas (la quinta subsección detalla cómo se construye la variable de habilidad), esta variable puede servir para confirmar o verificar la clasificación de vacantes cuando los títulos de trabajo no son suficientes (ambiguos) para determinar un código CUOC a nivel de 5 dígitos. El tercer ingrediente principal, proviene del marco ampliado CUOC; este archivo contiene el CIUO a nivel de 4 dígitos, también incluye el grupo (preliminar) de subunidades

(códigos) y el grupo de títulos (etiquetas) para los grupos ocupacionales a nivel de 5 dígitos. Esta última variable será particularmente útil para codificar los datos de vacantes, como se muestra en los siguientes apartados.

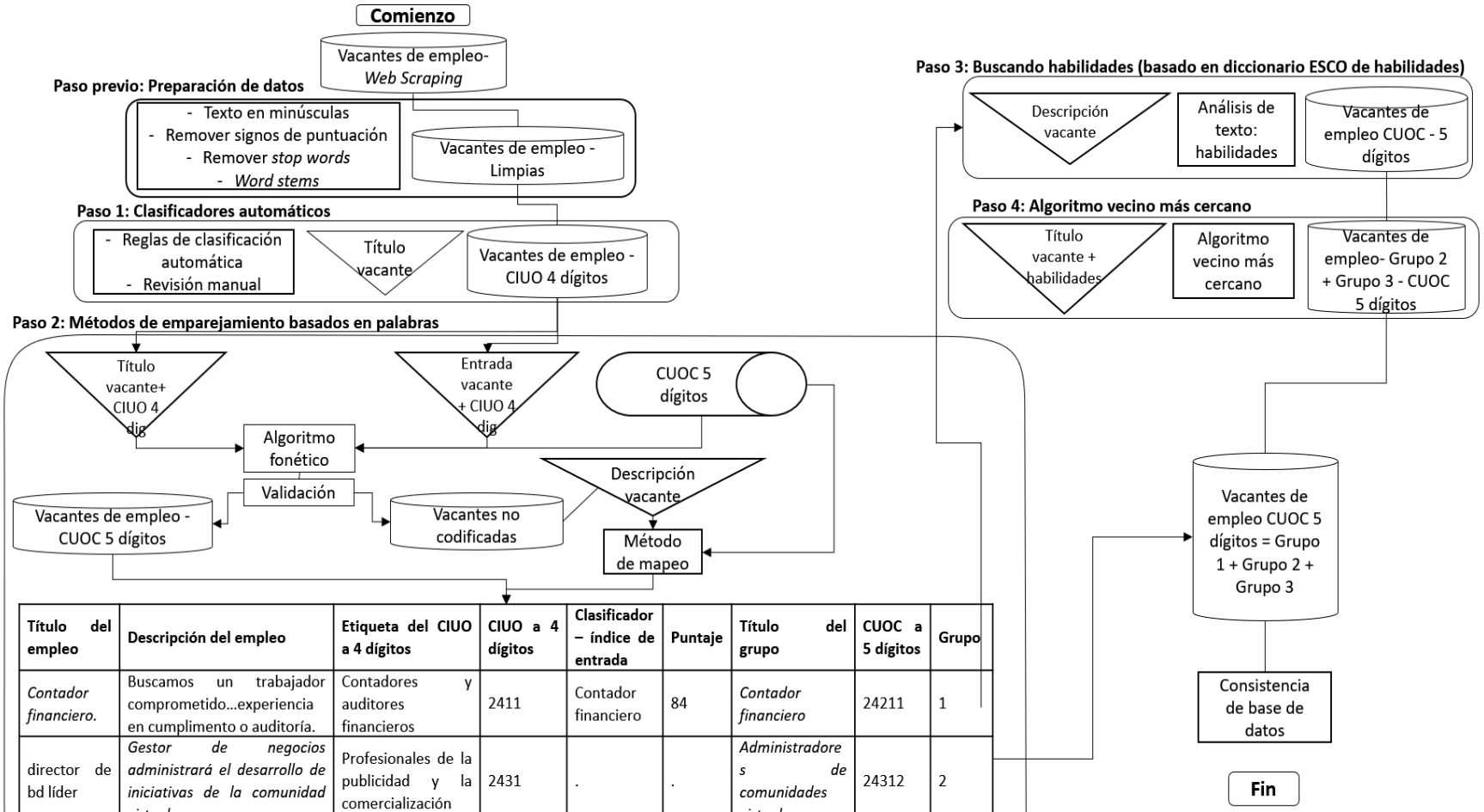
En términos de los métodos de clasificación, se propone utilizar los clasificadores automáticos utilizados por Cardenas (2020), el método de *emparejamiento basados en palabras* (también conocido como métodos "*fuzzy merge*") y algoritmos de *machine learning* (*algoritmo del vecino más cercano*). Estos métodos han sido probados en diferentes bases de datos para clasificar los títulos de trabajo en códigos ocupacionales con resultados de alta precisión (Jones y Elias, 2004; Gweon et al. 2017; Lima y Bakhshi, 2018).

La figura 1 muestra los pasos para codificar y probar la base de datos de vacantes a nivel de 5 dígitos de la CUOC. Esta metodología comienza con la base de datos de vacantes extraída de los portales de empleo mencionados previamente. Antes del procesamiento de los datos, se prepara la base de datos de vacantes para este análisis. Específicamente, este proceso implica la implementación de técnicas de *text mining* (tales como *stop words, stemming*, etc) para limpiar la información "ruidosa" (preposiciones, símbolos, etc.) Una vez depurados los datos, la metodología propuesta seguirá los siguientes cuatro pasos:

- Uso de clasificadores automáticos.
- Métodos de *emparejamiento basados en palabras*.
- Mapeo de habilidades en la base de datos de vacantes.
- Aplicación del algoritmo del vecino más cercano.
- Prueba de consistencia.

En las siguientes subsecciones explicaremos en detalle cada uno de los pasos para clasificar y probar la base de datos de vacantes a nivel de 5 dígitos de la CUOC.

Figura 2. Esquema general para codificar los datos de vacantes a nivel de 5 dígitos de la CUOC



Título del empleo	Descripción del empleo	Etiqueta del CIUO a 4 dígitos	CIUO a 4 dígitos	Clasificador – índice de entrada	Puntaje	Título del grupo	CUOC a 5 dígitos	Grupo
Contador financiero.	Buscamos un trabajador comprometido…experiencia en cumplimento o auditoría.	Contadores y auditores financieros	2411	Contador financiero	84	*Contador financiero*	24211	1
director de bd líder	*Gestor de negocios administrará el desarrollo de iniciativas de la comunidad virtual.*	Profesionales de la publicidad y la comercialización	2431	.	.	*Administradores de comunidades virtuales*	24312	2
Líder de finanzas operativas.	Ubicación financiera: Actualmente estamos trabajando de forma remota y te apoyaremos.	.	.	.	.	.	.	3

2.1.1. Clasificadores automáticos.

Los clasificadores automáticos utilizados en Cardenas (2020) son una herramienta de clasificación para asignar un código ocupacional (o industrial) a los textos. Durante las últimas dos décadas y los acuerdos de asociación con expertos del mercado laboral, estas herramientas han diseñado un conjunto de reglas probadas (como palabras degradadas, finales de palabras equivalentes, abreviaturas, palabras de remplazo, etc.) e índices de puestos para revelar las mejores coincidencias entre puestos de trabajo y clasificaciones ocupacionales (Jones & Elias, 2004; Wageindicator, 2009; IER, 2018). Dichas herramientas se actualizan y puede codificar títulos de trabajo para CIUO a nivel de 4 dígitos. Por lo tanto, el uso de ellas se considera un buen punto de partida para clasificar los puestos de trabajo vacantes al nivel de 4 dígitos del CIUO y, potencialmente, estas entradas podrían ayudar a desglosar la información de vacantes en 5 dígitos de la CUOC.

Por lo tanto, el primer paso de esta metodología es codificar la base de datos de vacantes al CIUO a nivel de 4 dígitos utilizando herramientas clasificadoras. La codificación procede de la siguiente forma: los títulos de trabajo limpios de la base de datos de vacantes son la variable de entrada que las herramientas utilizarán para asignar el CIUO a la mejor coincidencia de 4 dígitos con las puntuaciones de similitud correspondientes y la entrada de índice más cercana[7].

La tabla 1 muestra un ejemplo de los resultados de arrojados por las herramientas clasificadoras utilizadas en Cardenas (2020). La primera columna indica una lista de puestos de trabajo de la base de datos de vacantes, mientras que el resto de las columnas muestran los resultados después de ejecutar el proceso. La primera fila de esta tabla indica que una empresa necesitaba “dibujantes técnicos”. El proceso sugiere que la mejor coincidencia es “Dibujante técnico”. con una puntuación de similitud de 57[8]; el resultado indica que el código ocupacional de esta observación es “3120 CAD, Técnicas de dibujo y arquitectura”. De forma similar, la segunda y la tercera fila muestran otros

[7] Una lista complete de denominaciones de trabajo (o títulos de trabajo) asociados con un código ocupacional a nivel de cuatro dígitos.
[8] En una escala de 0 a 100, entre mas alto sea el número, mas alta cera el nivel de certeza de que el código encontrado es el correcto.

ejemplos. La última fila de la tabla 1, muestra un ejemplo donde el puntaje de similitud es relativamente bajo, como mencionan Jones & Elias (2004), un umbral razonable para aceptar el resultado de las herramientas es 40 o más. De esta forma, algunas observaciones en los datos de vacantes permanecerán sin codificar después de utilizar los clasificadores automáticos.

Tabla 1. Ejemplos de los resultados de herramientas clasificadoras

Título de la vacante (Obtenido de la base de vacantes)	**Entrada del índice.**	**Puntaje de similitud.**	**CIUO a 4-dígitos**
Dibujantes técnicos	Dibujante técnico	57	3120 CAD, Técnicas de dibujo y arquitectura
Coordinadora de necesidades educativas especiales (SENCO) primaria	SENCO	75	2316 profesores especializados en la educación especial
Ingeniero de sistemas	Ingenieros de sistemas	83	2133 IT Analistas de negocios, arquitectos diseñadores de sistemas.
Líder de finanzas operativas	Líder, equipo, operaciones, computadora	36	.

A pesar de que las herramientas de clasificadores automáticas presentadas por Cardenas (2020) proporcionan resultados precisos con un umbral de 40 o más, se seleccionará una muestra aleatoria de vacantes para verificar manualmente la precisión de la codificación. Después de la inspección visual, estos resultados se utilizarán en el siguiente paso de esta metodología para codificar la base de datos de vacantes la CUOC de 5 dígitos.

2.1.2. Métodos de Emparejamiento basados en palabras.

El segundo paso de esta metodología consiste en aplicar el método de *emparejamiento basados en palabras* entre la base de datos de vacantes y el CIUO. El método de *Emparejamiento basados en palabras* permite codificar de forma sencilla los datos de vacantes en el CUOC a nivel de 5 dígitos. En términos generales, estos métodos son un conjunto de algoritmos que comparan palabras y frases que coinciden, que están por encima de un determinado umbral de puntuación de coincidencia.

En este caso, los algoritmos podrían usar el título del trabajo o la entrada del índice (variables disponibles en la base de datos de vacantes) y relacionarlas con el título del grupo en la CUOC. El resultado de este proceso podría asignar una ocupación con un nivel del código a 5 dígitos para un conjunto de observaciones en la base de datos de vacantes[9].

Como se observa en la
Figura 2, existen dos formas de emparejar los datos de vacantes con el CUOC 5-digitos. La primera forma, corresponde a utilizar el título del trabajo y el CIUO de 4 dígitos como variables clave de la base de datos de vacantes y el título de grupo y las variables de CIUO de 4 dígitos. Utilizar la variable CIUO a nivel de 4 dígitos reduce el número de posibles coincidencias producidas por "Título del trabajo - Título del grupo" y, por lo tanto, podría ayudar a obtener resultados más precisos. Se esperaría que el método asigne un código de ocupación a nivel de 5 dígitos al conjunto de observaciones en la base de datos de vacantes con un nivel de precisión significativo.

La segunda forma es una combinación de las bases de datos de vacantes, el CIUO y la CUOC. Con lo anterior se utiliza el índice de entrada y las variables del CIUO a 4 dígitos de la base de datos de vacantes y el título del grupo, así como las variables CIUO a 4 dígitos. El uso de las variables CIUO a 4 dígitos reducen el número de posibles emparejamientos producidos por "Índice de entrada – Título del grupo", por lo tanto, podría ayudar a obtener resultados más precisos.

El emparejamiento de las bases de datos con los clasificadores automáticos de Cardenas (2020) podría tener dos propósitos: En primer lugar, podría ayudar a clasificar las observaciones no codificadas por "Título del trabajo -Título del grupo", de tal manera que podría existir el caso en el que el índice de entrada sería una mejor variable para emparejar los datos, en comparación con la variable de título de la vacante. En segundo lugar, este proceso podría utilizarse para validar, hasta cierto punto el emparejamiento por "Título del trabajo -Título del grupo". Los resultados equivalentes uniendo las

[9] Debido a que existen diferentes enfoques, cada uno con sus propias ventajas y desventajas, sería necesario combinar diferentes métodos de emparejamiento basados en palabras. Esta combinación se decidirá sobre la base de una revisión de literatura, así como de los resultados empíricos de la implementación de estos métodos en la base de vacantes. De esta forma, es necesario realizar una prueba de sensibilidad para garantizar un nivel de certeza relativamente alto en la codificación de la base de datos de vacantes.

bases de datos por “Título del trabajo -Título del grupo” e “Índice de entrada -Título del grupo”, pueden corresponder a un alto nivel de ajuste en la codificación de la base de datos de vacantes.

Por lo anterior, cuando existan discrepancias entre los procesos de combinación, se utiliza como resultado final aquel proceso que tenga un puntaje mayor en el método de emparejamiento basados en palabras.

Adicionalmente, se realizarán inspecciones visuales para garantizar en cierta medida, el nivel de precisión de los resultados y evaluar casos complejos (p. ej. Puntuaciones de similitud similares)
La tabla 2 muestra un ejemplo de la base de datos de vacantes después de aplicar los dos primeros pasos de esta metodología. Las dos primeras columnas de la tabla 2 muestran las variables que provienen de los portales de empleo (títulos del empleo, descripción del empleo, entre otras). Desde la columna tres hasta la seis, muestran la creación de variables después de aplicar los clasificadores automáticos. La columna séptima y octava (“Titulo del grupo” y “CUOC 5 dígitos”) indican las variables creadas después de aplicar el método de *emparejamiento basados en palabras*. La última variable (“Grupo”) sirve para explicar los siguientes pasos de esta metodología.

La variable grupo indica los diferentes grupos que podrían resultar después de aplicar los dos primeros pasos de esta metodología sobre la base de datos de vacantes. El grupo 1 corresponde a aquellas observaciones que se codificaron con los clasificadores automáticos y donde lo métodos de *emparejamiento basados en palabras* encuentran una mejor coincidencia de la CUOC (ya sea con las combinaciones “Título del trabajo -Título del grupo” o “Índice de entrada -Título del grupo”). Por lo tanto, este grupo de observación puede considerarse codificado a nivel de 5 dígitos de la CUOC. No obstante, se desconoce el número de posibles observaciones en este y en los otros grupos. El grupo 2, indica aquellas observaciones codificadas con las herramientas clasificadores, pero que no encuentran una mejor coincidencia de la CUOC utilizando los métodos de *emparejamiento basados en palabras*. Por último, el grupo 3 representan las observaciones no codificadas por las herramientas clasificadores ni por los métodos de *emparejamiento basados en palabras*. Las observaciones en el grupo 2 y en el grupo 3 requieren un mayor análisis para codificarlas de acuerdo con la CUOC a un nivel de 5 dígitos.

Tabla 2. Ejemplo de la base de datos de vacantes después de aplicar herramientas clasificadoras y los métodos de emparejamiento basados en palabras

Título del empleo	Descripción del empleo	Etiqueta del CIUO a 4 dígitos	CIUO a 4 dígitos	Clasificador – índice de entrada	Puntaje	Título del grupo	CUOC a 5 dígitos	Grupo
Contador financiero.	Buscamos un trabajador comprometido…experiencia en cumplimento o auditoría.	Contadores públicos y contadores certificados.	2421	Contador financiero	84	Contadores financieros	2421/02	1
Director de recaudación de fondos.	Gerente de recaudación de fondos. Bishops gate Institute es un hogar para las ideas.	Directores de marketing, ventas y publicidad.	1132	Director (publicidad)	93	Directores de publicidad	1132/01	1
Gerente de extracción de fósiles.	Como director de planificación, trabajará en estrecha colaboración con el director y otros miembros.	Gerentes y directores de producción en minería y energía.	1123	Gerente (canteras y extracción)	53	Gerentes y directores en la extracción de combustibles fósiles	1123/04	1
SENCO primaria.	Esta es una oportunidad emocionante para un experimentado.	Profesional de la enseñanza de educación especial	2316	SENCO	75	.	.	2
Líder de finanzas operativas.	Ubicación financiera: Dunstable, Bedfordshire. Actualmente estamos trabajando de forma remota y te apoyaremos.	.	.	.	.	.	.	3

2.1.3. Buscando habilidades.

Como se mencionó anteriormente los datos de vacantes podrían codificarse a nivel de 5 dígitos CUOC porque estas fuentes contienen una serie de variables detalladas que pueden servir como insumos relevantes en el proceso de codificación. Los pasos anteriores utilizan la variable del título del trabajo para realizar un primer enfoque de codificación a nivel de 5 dígitos. A pesar de lo anterior, se necesita información más detallada para codificar el conjunto de observaciones que no fueron codificados en el primer paso de la metodología.

En específico, las descripciones de puestos vacantes pueden usarse para ayudar a codificar esas observaciones "difíciles de codificar". La descripción del puesto es un texto abierto donde los empleadores detallan las características de los candidatos requeridos. Estas variables es un insumo importante para identificar los requisitos del mercado laboral, como calificaciones, experiencia y requisitos de habilidades, etc. Los requisitos de habilidades pueden ser información crucial para codificar la base de datos de vacantes a nivel de 5 dígitos CUOC. Autores como Lima & Bakhshi (2018) y Cárdenas-Rubio (2020) han utilizado los requisitos de habilidades como entrada para asignar códigos ocupacionales (nivel de 4 dígitos) a los datos de vacantes con un alto nivel de precisión.

En consecuencia, esta metodología propone explotar la posibilidad de codificar esas observaciones "difíciles de codificar" en la base de vacantes de Colombia en un nivel de 5 dígitos con la ayuda de los requisitos de habilidades. Para hacer esto, primero, es necesario identificar las habilidades que demandan los empleadores. Como se mencionó previamente, los empleadores usan la descripción del trabajo para enumerar las habilidades necesarias para un determinado trabajo. Sin embargo, los requisitos de habilidades no suelen estar organizados en variables separadas ni categorizadas bajo la misma tipología (Cárdenas-Rubio, 2020). Este problema dificulta el uso de la información sobre habilidades para el análisis estadístico. Por lo tanto, es necesario realizar técnicas de extracción de textos sobre los datos de las vacantes para obtener una lista completa y estandarizada de las habilidades requeridas para cada trabajo.

Se podrían utilizar diferentes enfoques para estandarizar la información sobre habilidades. Por ejemplo, Burning Glass usa su diccionario de habilidades para analizar los patrones de habilidades de las descripciones de puestos y agrupar esta información en diferentes categorías. Asimismo, el "Skills-

OVATE: Skills Online Vacancy Analysis Tool for Europe" (Cedefop, 2019) y Cárdenas-Rubio (2020) han utilizado las Habilidades, Competencias, Cualificaciones y Ocupaciones Europeas (ESCO[10] por sus siglas en inglés) para identificar patrones de habilidades en datos de vacantes europeas y latinoamericanas. Dado que la ESCO es un diccionario público, esta metodología utilizará el diccionario de la ESCO para identificar y estandarizar la información de habilidades en la base de datos de vacantes.

Como explica Cárdenas-Rubio (2020), la identificación de patrones de habilidades con el diccionario ESCO y los datos de vacantes se desarrolla de la siguiente manera:

- Las palabras comunes (como preposiciones, palabras vacías) se excluyen desde el diccionario ESCO y de la descripción del trabajo en el conjunto de datos vacantes.
- Las letras se transforman a minúsculas y las palabras se reducen a su raíz gramatical tanto en el diccionario ESCO como en la descripción de la base de datos de vacantes.
- Cada palabra o frase en el diccionario de habilidades se busca en cada anuncio de vacante de empleo. Esta exploración de palabras se codifica en variables de unigrama (n-grama), que son variables indicadoras. Las variables toman el valor 1 si una determinada palabra o frase (patrón) en el diccionario de habilidades se encuentra en un anuncio y 0 en caso contrario (se crea un máximo de 13,485 variables)
- Es necesario realizar verificaciones manuales para garantizar un cierto nivel de precisión del mapeo de habilidades.

Con estos procedimientos, la información sobre habilidades se clasificará con la base de datos de vacantes de Colombia. Es importante señalar que este método de mapeo tiene varias limitaciones (p. ej. no identifica sinónimos, los empleadores no siempre mencionan los requisitos de habilidades explícitamente en los anuncios de trabajo, etc.) A pesar de las limitaciones, el mapeo de vacantes ESCO podría proporcionar nuevos insumos (patrones) para clasificar las vacantes laborales en el nivel de 5 dígitos de la CUOC.

[10] La ESCO proporciona una lista completa de 13,485 habilidades utilizadas en el mercado laboral.

2.1.4. Algoritmo del vecino más cercano.

Los métodos de *machine learning*[11] se han utilizado recientemente como una alternativa para codificar títulos de trabajo en ocupaciones a nivel de 4 dígitos (Gweon at al. 2017; Lima y Bakhshi, 2018; Cárdenas-Rubio, 2020). En general, estos métodos, están compuestos por 1) una base de datos precodificada que sirve como ejemplo (base de datos de aprendizaje) para calibrar el modelo de aprendizaje automático y 2) un conjunto de reglas de asociación (algoritmo) para vincular títulos de trabajo con ocupaciones. Con estos dos elementos, el computador "aprende" reglas de asociación para codificar títulos de trabajo en ocupaciones.

Estos métodos podrían potencialmente usarse para ayudar a clasificar la base de datos de vacantes en CUOC a nivel de 5 dígitos. Como se muestra en la tabla 2, las observaciones de vacantes se pueden clasificar en tres grupos. El grupo 1 corresponde a todas las observaciones codificadas a nivel de 5 dígitos utilizando los clasificadores automáticos presentadas por Cardenas (2020), métodos de *emparejamiento basados en palabras* e inspecciones visuales. Este conjunto de observaciones podría usarse como la base de datos precodificada que sirve como ejemplo para calibrar el modelo de *machine learning.*

Con respecto al algoritmo de *machine learning*, existen diferentes métodos (p. ej. *Support Vector Machine, Random Forest*, etc.) que se pueden usar para clasificar los títulos de trabajo en ocupaciones. La mayoría de los métodos de *machine learning* utilizan los títulos de trabajo para aprender a codificar observaciones en ocupaciones. Sin embargo, Lima & Bakhsi, 2018 y Cárdenas-Rubio (2020) han demostrado que las variables de habilidades podrían usarse junto con los títulos de trabajo para mejorar considerablemente los resultados de los algoritmos de *machine learning*. En particular, Cárdenas-Rubio (2020) encuentra que el algoritmo del vecino más cercanoproporciona resultados consistentes a nivel de 4 dígitos utilizando tanto los títulos de trabajo como la información de habilidades. Además, este algoritmo es relativamente fácil de interpretar y no requiere una capacidad computacional extremadamente alta.

[11] *Machine Learning* puede definirse como: "conjunto de métodos que pueden detectar automáticamente patrones en los datos y luego usar los patrones descubiertos para predecir datos futuros o para realizar otros tipos de toma de decisiones en condiciones de incertidumbre" (Murphy, 2012, p.1)

En términos generales, este algoritmo podría utilizar los títulos de trabajo, las variables de habilidad (de la base de datos precodificada) y una versión mejorada del puntaje de similitud de coseno para determinar qué patrones podrían indicar un determinado grupo ocupacional (el vecino más cercano) (Ver Cárdenas-Rubio (2020), y Gweon el al. 2017 para más detalles).

En el caso particular de codificar la base de vacantes al nivel de 5 dígitos de la CUOC, el algoritmo del vecino más cercano podría implementarse usando:

- El diccionario de competencias de la ESCO y,
- La variable "Título del grupo" del marco CUOC.

2.1.5. Implementando el algoritmo del vecino más cercano en la base de datos de vacantes.

Como se mencionó anteriormente, las observaciones en el grupo 1 (ver tabla 2) en los datos de vacantes pondrían usarse como base de datos de aprendizaje. Estos datos se dividirán en las bases de datos de entretenimiento y prueba, para calibrar los parámetros del algoritmo del vecino más cercano. Adicionalmente, como se muestra en la tabla 2, las observaciones que quedaron sin codificar están compuestas por aquellas observaciones que fueron codificadas con clasificadores automáticos a nivel de cuatro dígitos, pero que no fueron codificadas exitosamente con el método de emparejamiento basado en palabras a nivel de 5 dígitos (grupo 2) y aquellas observaciones que los clasificadores automáticos y los métodos de emparejamiento basado en palabras no pudieron codificarlos a nivel de 4 dígitos del CIUO.

El grupo 2 y grupo 3 se pueden utilizar por separado para identificar su código ocupacional a nivel de 5 dígitos. El grupo 2 se caracteriza por estar codificado a nivel de 4 dígitos del CIUO. Esta variable se puede utilizar como insumo en el algoritmo del vecino más cercano para garantizar resultados con precisión relativamente alta. La tabla 3 muestra un ejemplo simple de cómo este algoritmo podría funcionar en la práctica.

Supongamos que tenemos una base de datos precodificada (de entrenamiento) con cuatro observaciones: Las cuatro observaciones en estos datos exigen un gastroenterólogo, un cardiólogo y

un dermatólogo (las primeras cuatro filas de la tabla 3). Estas observaciones están codificadas a nivel de 4 dígitos del CIUO como “2212 Médicos especialistas”. Dados los títulos de los empleos es posible conocer los códigos ocupacionales a nivel de 5 dígitos (última columna de la tabla 3).

Ahora bien, supongamos además que existe una observación no codificada en el grupo 2 que exige un médico especialista. Esta observación se codifica como “2212 Médicos especialistas”. Sin embargo, es difícil determinar el nivel de 5 dígitos de esta observación considerando solo el puesto de trabajo (última fila de la tabla 3). En la CUOC, el título de trabajo del médico especialista podría codificarse como gastroenterólogo, cardiólogo o dermatólogo, etc. Por lo tanto, la información sobre habilidades podría usarse para determinar precisamente este quinto nivel. Las columnas 7 a 10 muestran un ejemplo de variables creadas después de mapear el diccionario ESCO en la base de vacantes. Se identificaron tres habilidades: gastroenterología, cardiología y dermatología.

En este caso, la observación no codificada en el grupo 2 requiere conocimientos en gastroenterología. El vecino más cercano de esta observación dentro de los “2212 médicos especialistas” en el conjunto de datos precodificados (grupo 1) es “Gastroenterólogo” (primera fila de la tabla 3), dado que comparten el mismo conjunto de habilidades exigidas o las más similares. En consecuencia, la observación del grupo 2 se codificará como “2212/04 Gastroenterólogos”. Es importante señalar que este ejemplo no considera escenarios donde, por ejemplo, hay más de un vecino más cercano. Estos casos se evaluarán a medida que se procesen los datos.

Tabla 3. Ejemplo del algoritmo del vecino más cercano utilizando código y habilidades del CIUO a 4 dígitos.

Fuente	Título de empleo						Habilidades			CIUO a 4 dígitos	CUOC a 5 dígitos
	Gastroenterólogo	Cardiólogo	Dermatólogo	Especialista	Médico	Facultativo	Gastroenterología	Cardiología	Dermatología		
Datos de entrenamiento (Grupo 1)	1	0	0	0	0	0	1	0	0	2212 Médicos especialistas	2212/04 Gastroenterólogos
	0	1	0	0	0	0	0	1	0	2212 Médicos especialistas	2212/02 Cardiólogos
	0	0	1	0	0	0	0	0	1	2212 Médicos especialistas	2212/03 Dermatólogos
	0	0	1	0	0	0	0	0	1	2212 Médicos especialistas	2212/03 Dermatólogos
Resultado (Grupo 2)	0	0	0	1	1	1	1	0	0	2212 Médicos especialistas	2212/04 Gastroenterólogos*

A la luz de lo anterior, se podría emprender un enfoque similar para codificar las observaciones en el grupo 3. No obstante, para este grupo, el CIUO a nivel de 4 dígitos no está disponible. En este caso, será necesario utilizar los títulos de trabajo (en lugar de los códigos CIUO) y la información de habilidades para determinar el vecino más cercano. La tabla 4 muestra un ejemplo reducido de cómo el algoritmo del vecino más cercano podría funcionar en la práctica. Por un lado, supongamos que tenemos una base de datos de entrenamiento. En esta base de datos, se demandan y codifican cuatro puestos de trabajo: Diseñador web (código CIUO 2141/03 Diseñadores web); Diseñador gráfico (2132/99 Diseñadores gráficos y multimedia n.e.p) y Diseñadores de joyas (3422/03 Diseñadores de joyas). Por otro lado, hay una observación en el grupo 3 cuyo puesto de trabajo es "diseñador". Con solo esta entrada, no es posible determinar el código CIUO y CUOC en niveles de 4 y 5 dígitos. Sin embargo, la información de habilidades de la descripción de la vacante de trabajo podría proporcionar información adicional para determinar los códigos ocupacionales correspondientes. Las columnas 6 a 8 indican las habilidades requeridas. Como se puede observar, el vecino más cercano a la observación del grupo 3 es el "Diseñador gráfico" (tercera fila de la tabla 4) ya que comparte más habilidades similares que las otras observaciones en la base de datos de capacitación. Así, se asignará a la observación del grupo 3 el código ocupacional a nivel de 5 dígitos "2142/99 Diseñadores gráficos y multimedia n.e.c. *". Como se mencionó anteriormente, es importante tener en cuenta que este ejemplo no considera escenarios donde, por ejemplo, hay más de un vecino más cercano. Estos casos se evaluarán a medida que procesen los datos.

Alternativamente, en lugar de usar el diccionario ESCO para identificar las habilidades demandadas y usarlas como entrada para codificar las observaciones de vacantes al nivel de 5 dígitos de la CUOC, es posible usar la variable "Título del grupo" (etiqueta) del marco ampliado de la CUOC como un diccionario. Estas palabras se pueden asignar a los vecinos. En esta etapa de la propuesta, es imposible determinar qué opción; las habilidades de la ESCO o la variable "Título del grupo" del marco ampliado de la CUOC, podrían proporcionar mejores resultados. Por ahora, vemos estas opciones como alternativas complementarias y ambas deberían explorarse.

Con la implementación de los pasos anteriores, esperamos tener un número de observaciones de vacantes codificadas en 5 dígitos de la CUOC. Sin embargo, en esta etapa no es posible determinar la fracción de la base de datos de vacantes que permanecerá sin codificar. La clasificación de estos valores faltantes podría ser parte de un trabajo futuro.

Tabla 4. Ejemplo del algoritmo del vecino más cercano utilizando títulos de empleo y habilidad.

Fuente	Título de empleo				Habilidades			CIUO 4 a dígitos	CUOC a 5 dígitos
	Web	Gráfico	Joyería	Diseñador	ASP.NET	Pintura 2D	Cortar piedras preciosas		
Datos de entrenamiento (Grupo 1)	1	0	0	1	1	0	0	2141 Profesionales del diseño web.	2141/03 Diseñadores web
	0	1	0	1	0	1	0	2142 Diseñadores gráficos y multimedia.	2142/99 Diseñadores gráficos y multimedia n.e.c.
	0	0	1	1	0	0	1	3422 Diseñadores de ropa, moda y complementos.	3422/03 Diseñadores de joyas.
	0	0	1	1	0	0	1	3422 Diseñadores de ropa, moda y complementos	3422/03 Diseñadores de joyas.
Resultado (Grupo 3)	0	0	0	1	0	1	0	2142 Diseñadores gráficos y multimedia	2142/99 Diseñadores gráficos y multimedia n.e.c.*

3. Probar la coherencia de la base de datos de vacantes codificada a nivel de 5 dígitos de la CUOC.

El último paso de esta metodología consiste en realizar una primera evaluación experimental de la consistencia de la base de datos de vacantes a nivel de 5 dígitos de la CUOC. Los pasos anteriores han demostrado que una serie de observaciones de vacantes se pueden codificar a nivel CUOC de 5 dígitos. Sin embargo, aún se debe comprobar si los datos de vacantes pueden ser utilizados para la inferencia estadística de la demanda laboral insatisfecha en Colombia.

Como mencionan Cárdenas-Rubio (2020) y Kureková et al. (2014), es difícil probar la representatividad de los datos de vacantes ya que (generalmente) los países no tienen un censo de vacantes que proporcione el número total de vacantes (el universo estadístico) que se puede utilizar como punto de comparación. Cárdenas-Rubio (2020) y Štefánik (2012) han realizado pruebas de correlación con datos de oferta laboral para brindar información sobre la representatividad de los portales de empleo en línea. A pesar de las limitaciones de las pruebas, estos autores concluyeron que los portales de empleo en línea tienden a representar relativamente bien la dinámica laboral de un conjunto considerable de trabajadores formales, no agrícolas, no gubernamentales, no militares y no autoempleados (“propietarios de empresas”). Sin embargo, no es posible (en este momento) inferir que el número de vacantes de trabajo publicadas en línea corresponda al número total de vacantes disponibles en una economía, principalmente, debido a la ausencia de un censo de vacantes.

Además, estas pruebas se han realizado a nivel de 4 dígitos de la CIUO. Hasta donde sabemos, no se han realizado pruebas representativas a un nivel inferior. Probar la representatividad de los datos a un nivel más desagregado es aún más desafiante, ya que también es necesario tener un censo de vacantes desagregado en un nivel más bajo (es decir, quinto nivel). Por lo anterior, el último paso de esta metodología se centrará en probar la volatilidad de los datos de vacantes en lugar de la representatividad de estos datos.

La volatilidad de los datos se refiere a la tasa de cambio en el número de vacantes (al nivel de 5 dígitos de la CUOC) publicadas durante un período. La volatilidad de los datos de vacantes podría determinar los grupos ocupacionales, que muestran un número relativamente estable de vacantes publicadas a lo largo del tiempo. Estos resultados proporcionan una indicación de que los datos de esos grupos "estables" podrían usarse para analizar las tendencias de vacantes a mediano y (eventualmente) a largo plazo. Por ejemplo, una serie de vacantes "estable" a nivel de 5 dígitos con una tendencia positiva podría indicar que el tipo de trabajo es cada vez más demandado. Las series ocupacionales altamente volátiles sugerirían que los datos no podrían usarse para analizar las tendencias del mercado laboral. Existen diferentes enfoques para medir la volatilidad de los datos (p.ej. desviaciones estándar, varianza, modelos ARCH, etc). Sin embargo, los detalles sobre qué método se adapta mejor a la base de datos de vacantes dependerán de los resultados de los pasos anteriores. Parámetros a considerar al determinar la volatilidad:¨

- La proporción de missing en la variable de 5 dígitos de la CUOC.
- El número de observaciones por grupos ocupacionales. Debido a la estructura del mercado laboral de Colombia, algunos grupos ocupacionales tienen una mayor participación en el mercado laboral que otros (por ejemplo, ocupaciones agrícolas frente a ocupaciones administrativas). Las ocupaciones de baja frecuencia pueden tener un coeficiente de volatilidad más alto porque tienen una baja participación en el mercado laboral en lugar de que la base de datos de vacantes no recopile suficiente información. En consecuencia, el indicador de volatilidad debe considerar el tamaño del mercado laboral por grupos ocupacionales. Una forma de conocer estos tamaños es analizando la proporción de empleo por ocupaciones (acciones del mercado laboral). Se supone que un mayor stock laboral se correlaciona con un mayor número de vacantes (Cárdenas-Rubio, 2020). Por lo tanto, estos resultados indican las ocupaciones que podrían tener una proporción mayor/menor de vacantes dada la estructura del mercado laboral en Colombia.
- El tiempo. Como se mencionó en la introducción, Cardenas (2020) ha extraído datos de los portales de empleo desde 2018. Sin embargo, los efectos de Covid-19 en marzo de 2020 podrían haber afectado la estabilidad de la serie temporal. Se realizará un análisis para determinar el período a considerar para las estimaciones de volatilidad.

Este análisis muestra la fracción de la base de datos de vacantes que podría codificarse en 5 dígitos de la CUOC. Los resultados revelan que los títulos de empleo que son relativamente fáciles/difíciles de codificar; proporcionan una lista de títulos de empleo que no se pueden clasificar en un nivel de 5 dígitos, y podría sugerir nuevas vías de investigación para codificar esas observaciones. Además, se desarrolla una regla de decisión (coeficiente de volatilidad) para determinar si los datos se pueden utilizar a un nivel de 5 dígitos para la inferencia estadística (análisis de tendencias). Por último, los resultados de este proyecto podrían utilizarse para abrir nuevas vías de investigación.

3.1. Tendencia de ocupados entre 2008 y 2019

Este indicador busca describir la tendencia que ha mostrado el nivel de ocupados durante los años 2008 a 2019 para las ocupaciones a 5 dígitos sobre las cuales se tiene información disponible con la calidad admisible. La categorización se califica de acuerdo con el grado de significancia estadística y de nivel de que tengan los cambios de la serie de cada ocupación a lo largo del tiempo a apartir de una regresión con la variable de tiempo.

Se trata de una categorización de las tendencias de las ocupaciones en la siguiente forma:

Tendencia	Descripción
Creciente fuerte	Crecimientos sostenidos de tendencia pronunciada con significancia estadística
Creciente moderado	Crecimientos bajos con una significancia baja estadísticamente.
Estable	No ostenta cambios significativos en la tendencia y los cambios no tienen significancia estadística.
Decreciente moderado	Reducciones sostenidas con una significancia baja.
Decreciente fuerte	Reducciones de tendencia pronunciada con significancia estadística
No disponible	Sin información disponible

3.2. Número de vacantes disponibles

Corresponde al número de vacantes registradas en la base del Servicio público de empleo que se ha logrado asociar a cada una de las ocupaciones definidas en la CUOC.

3.3. Rango de salarios estimado para los ocupados entre 2017 y 2019 en términos reales de 20018 (deflactados por IPC=2018)

El rango de salarios corresponde al rango que cada ocupación tuvo en el promedio de los ingresos entre los años 2017 y 2019 a pesos constantes de 2018. Este rango permite evidenciar el mínimo y máximo de salarios percibidos por la población ocupada en cada una de las ocupaciones de la CUOC.

3.4. Tablero estadístico

El tablero estadístico del Catálogo de Ocupaciones para Colombia -Ocupacol- es una herramienta que presenta los principales indicadores del mercado laboral a nivel de ocupación provenientes de la Gran Encuesta Integrada de Hogares -GEIH- del DANE y del Servicio Público de Empleo -SPE-.

A continuación, se presenta el listado, la descripción y el cálculo utilizada para todos los indicadores presentes en el Tablero Estadístico, cada indicador se calcula sobre un contexto en específico, es decir se filtra la información para no duplicar los datos.

Sección	Indicador	Descripción	Cálculo	Filtro o Contexto
Resumen	Ocupaciones	Cantidad de Ocupaciones en la CUOC	Conteo Distinto (Código Cuoc)	Total de ocupaciones de la CUOC
Resumen	Denominaciones	Cantidad de Ocupaciones en la CUOC	Conteo Distinto (Código Denominaciones)	Total de Denominaciones de la CUOC
Resumen	Total de ocupados	Número de personas ocupadas según la GEIH	Suma (Número de Ocupados)	Sobre la información del archivo número de ocupados
Resumen	Total de desocupados	Número de personas desocupadas según la GEIH	Suma (Número de Desocupados)	Sobre la información del archivo número de desocupados
Resumen	Total de Ofertas laborales	Número de vacantes en un periodo según el SPE	Suma (Número de Ofertas laborales)	Sobre la información del archivo número de vacantes
Salarios	Mediana de Salario	salario medio de la ocupación según encuestas de hogares (GEIH)	Mediana de salario* *Se presenta la mediana de salario por ocupación y por periodo de referencia y como mecanismo de agregación se utiliza el promedio de la mediana del salario en casos donde se tiene más de una ocupación	Sobre la información del archivo salarios
Salarios	Variación Mediana de Salario	Variación anual del salario medio con respecto al año anterior según encuestas de hogares (GEIH)	$\frac{M_e\ (S_n) - M_e(S_{n-1})}{M_e(S_{n-1})}\ x\ 100\%$ $M_e = Mediana$ $S_n = Salario\ Periodo\ Actual$ $S_{n-1} = Salario\ Periodo\ Anterior$	Sobre la información del archivo salarios sin tener en cuenta la información de las ocupaciones no clasificadas
Desocupados	Número de Desocupados	Número de personas desocupadas según	Suma (Número desocupados)	Sobre la información del archivo desocupados

		encuestas de hogares (GEIH)		
Desocupados	Variación desocupados	Variación anual del número de desocupados con respecto al año anterior según encuestas de hogares (GEIH)	$\frac{Sum\ (D_n) - Sum(D_{n-1})}{Sum(S_{n-1})}\ x\ 100\%$ $D_n = Desocupados\ Periodo\ Actual$ $D_{n-1} = Desocupados\ Periodo\ Acnterior$	Sobre la información del archivo desocupados sin tener en cuenta la información de las ocupaciones no clasificadas
Ocupados	Número de Ocupados	Número de personas ocupadas según encuestas de hogares (GEIH)	Suma (Número ocupados)	Sobre la información del archivo ocupados
Ocupados	Variación de ocupados	Variación anual del número de ocupados con respecto al año anterior según encuestas de hogares (GEIH)	$\frac{Sum\ (O_n) - Sum(O_{n-1})}{Sum(O_{n-1})}\ x\ 100\%$ $O_n = Ocupados\ Periodo\ Actual$ $O_{n-1} = Ocupados\ Periodo\ Acnterior$	Sobre la información del archivo ocupados sin tener en cuenta la información de las ocupaciones no clasificadas
Ofertas Laborales	Número de Ofertas laborales	Número de personas ocupadas según encuestas de hogares (GEIH)	Suma (Número ocupados)	Sobre la información del archivo vacantes
Ofertas Laborales	Variación de ofertas laborales	Variación anual del número de ofertas laborales con respecto al año anterior según el Servicio Público de Empleo	$\frac{Sum\ (Ol_n) - Sum(Ol_{n-1})}{Sum(O_{n-1})}\ x\ 100\%$ $Ol_n = Ofertas\ laborales\ Periodo\ Actual$ $Ol_{n-1} = Ofertas\ laborales\ Periodo\ Acnterior$	Sobre la información del archivo de vacantes sin tener en cuenta la información de las ocupaciones no clasificadas

Referencias

Cardenas Rubio, J. (2020). A web-based approach to measure skill mismatches and skills profiles for a developing country: the case of Colombia (Doctoral dissertation, University of Warwick).

Cedefop. 2019. Online Job Vacancies and Skills Analysis: A Cedefop Pan-European Approach. Luxembourg: Office for Official Publications of the European Communities.

GSS - The Government Statistical Service. 2021. "Standard Occupational Classifications (SOC) Extension Project" Retrieved April 10, 2020 (https://gss.civilservice.gov.uk/user-facing-pages/standard-occupational-classifications-soc/).

Gweon, Hyukjun, Matthias Schonlau, Lars Kaczmirek, Michael Blohm, and Stefan Steiner. 2017. "Three Methods for Occupation Coding Based on Statistical Learning." *Journal of Official Statistics* 33(1):101–22.

IER. 2018. "Cascot International." Coventry: University of Warwick. Retrieved September 20, 2018 (https://warwick.ac.uk/fac/soc/ier/software/cascot/internat/).

Jones, R., and P. Elias. 2004. CASCOT: Computer-Assisted Structured Coding Tool. Coventry: University of Warwick.

Dierdorff, E.C., Norton, J.J., Drewes, D.W., Kroustalis, C.M., Rivkin, D. & Lewis, P. (2009) Greening of the world of work: Implications for O* NET®-SOC and new and emerging occupations. National Center for O*NET Development, O*NET Resource Center (onetcenter.org).

Kanders, K, Djumalieva, J, Sleeman, C & Orlik, J 2020, Mapping career causeways: supporting workers at risk: a new system for supporting job transitions and informing skills policy in a changing labour market, Nesta, London, viewed 01 Mar 2021, <https://www.nesta.org.uk/report/mapping-career-causeways-supporting-workers-risk/>.

Kilhoffer, Z. (2020), Report on how to identify and compare newly emerging occupations and their skill requirements, Deliverable 12.2, Leuven, InGRID-2 project 730998 – H2020

Kureková, Lucia Mytna, Miroslav Beblavy, and Anna-Elisabeth Thum. 2014. *Using Internet Data to Analyse the Labour Market: A Methodological Enquiry*. IZA Discussion Paper 8555. Bonn, Germany.

Lewis, P., & Norton, J. (2018). Identification of "hot technologies" within the O* NET system.

Lima, Antonio, and Hasan Bakhshi. 2018. *Classifying Occupations Using Web-Based Job Advertisements: An Application to STEM and Creative Occupations*. ESCoE Discussion Paper 2018-08. London: Economic Statistics Centre of Excellence.

LinkedIn. 2021a. "career-explorer". Retrieved February 25, 2021 (https://linkedin.github.io/career-explorer/).

ONS. 2021a. "SOC 2020 Volume 1: structure and descriptions of unit groups" Retrieved April 10, 2020 (https://www.ons.gov.uk/methodology/classificationsandstandards/standardoccupationalclassificationsoc/soc2020/soc2020volume1structureanddescriptionsofunitgroups).

ONS. 2021b. "Classifying the Standard Occupational Classification 2020 (SOC 2020) to the International Standard Classification of Occupations (ISCO-08)" Retrieved April 10, 2020 (https://www.ons.gov.uk/methodology/classificationsandstandards/standardoccupationalclassificationsoc/soc2020/classifyingthestandardoccupationalclassification2020soc2020totheinternationalstandardclassificationofoccupationsisco08).

Sally-Anne Barnes, Bimrose Jenny, Cardenas-Rubio Jeisson, David Owen, Deirdre Hughes, Robert A. Wilson, Graham Attwell, and Philipp Rustemeier. *Labour market information (LMI) for all: Stakeholder Engagement and Usage, Data and Technical Developments*. Department for Education, 2021.

Sofroniou, N. & Anderson, P. (2021 forthcoming) 'The green factor: Unpacking green job growth', International Labour Review, doi.org/10.1111/ilr.12176, https://onlinelibrary.wiley.com/doi/full/10.1111/ilr.12176

Štefánik, Miroslav. 2012. "Internet Job Search Data as a Possible Source of Information on Skills Demand (with Results for Slovak University Graduates)." *Building on Skills Forecasts—Comparing Methods and Applications* 246.

Wageindicator. 2009. *EurOccupations: CASCOT Software for Coding Job Titles*.