I. Metodología

La publicación del Directorio de Empresas y Establecimientos de Cantabria es el resultado de la difusión de la operación estadística número 17.01 recogida en la **Ley de Cantabria 5/2016**, de 19 de diciembre, del Plan Estadístico 2017-2020.

El Directorio de Empresas y Establecimientos de Cantabria contiene los datos básicos de nombre, localización, empleo y tipo de sociedad del conjunto de las empresas y centros de producción de bienes o servicios que operan en Cantabria. Ambos aparecen clasificados en los términos estadísticos normalizados a nivel nacional e internacional: Clasificación Nacional de Actividades Económicas 2009 y NACE rev. 2.

Las **estadísticas sobre demografía de empresas** en Europa cubren todas las actividades excepto agricultura y pesca, para cualquier empresa de por lo menos un empleado, con referencia a las **unidades locales** (**establecimientos**) y a las actividades de acuerdo con la nomenclatura CNAE 2009.

Las **estadísticas estructurales sobre empresas** no son estadísticas sobre empresas, sino que describen los fenómenos económicos a través de la observación de las **unidades de actividad** implicadas en la actividad económica.

Finalidades

El Directorio de Empresas y Establecimientos Económicos sigue las recomendaciones internacionales en la materia. Las finalidades de esta operación estadística son las siguientes:

1) Detección y construcción de las unidades estadísticas existentes en nuestro ámbito territorial para cada momento de tiempo.
2) Instrumento para preparación y coordinación de encuestas.
3) Instrumento de movilización de datos administrativos para fines estadísticos.
4) Difusión.

El Directorio debe conformarse para responder a la necesidad de obtener información o marco para el sistema de estadísticas estructurales de empresas o cualquier operación de investigación estadística sobre empresas o sectores económicos.

El Directorio nos proporciona también un marco poblacional para ponderar la información de nivel inferior a las ramas de actividad que vayamos obteniendo.

Además, el **Directorio de Empresas y Establecimientos** es la fuente en que referir las estimaciones de empleo y asalariados.

El Directorio debe satisfacer la necesidad de comparabilidad entre las distintas estadísticas no sólo nacionales, sino también comunitarias, y por lo tanto, se hace necesario adoptar las definiciones y descriptores comunes del campo de las empresas y de las demás unidades cuya actividad es objeto de estadísticas.

Dichas definiciones y descriptores se describen en los reglamentos del Consejo de las Comunidades Europeas nº 177/2008 y nº 1893/2006 (que establece la nomenclatura estadística de actividades económicas de la Unión Europea -NACE Rev.2). El primero de dichos reglamentos establece un programa de armonización de estadísticas económicas promoviendo el desarrollo de directorios en los países miembros.

El Directorio produce dos tipos de productos que se suministran a los usuarios: relaciones pormenorizadas de unidades y tabulaciones estadísticas.

Las relaciones pormenorizadas de unidades (microdatos) que se facilitan son aquellas que permite la ***LEY 4/2005, de 5 de octubre, de Estadística de Cantabria****, que en el Artículo 38 excluye del secreto estadístico a: " Los directorios de establecimientos, empresas, organismos o entes de cualquier clase que no contengan más datos que la denominación, identificadores, emplazamiento, indicadores de actividad, tamaño y otras características generales que habitualmente se incluyan en los registros o directorios cuya publicidad haya sido declarada en normas con rango de Ley.* Por otra parte, el Reglamento europeo 223/2009 relativo a la estadística europea establece la necesidad de establecer principios y orientaciones comunes que garanticen la confidencialidad de los datos utilizados para elaborar estadísticas europeas y el acceso a esos datos confidenciales, habida cuenta del progreso técnico y de las necesidades de los usuarios en una sociedad democrática

La difusión de los microdatos se realiza a través de la siguiente dirección de la página web del ICANE: http://www.icane.es. Las tabulaciones estadísticas derivadas del directorio son las que se incluyen en esta publicación. A la hora de analizar unas y otras, hay que tener presente que el ámbito de difusión de microdatos está restringido a empresas y organismos, excluyendo por tanto las personas físicas. Las tabulaciones estadísticas se extienden a este colectivo. No obstante con el fin de preservar el secreto estadístico se excluyen las referencias a personal empleado cuando se tabula con un número inferior a tres registros.

Las tabulaciones estadísticas (o macrodatos de resultados) son las que se derivan o se pueden derivar de las variables y estados de las variables definidas para cada una de las unidades de referencia.

En las tabulaciones por actividad principal y ubicación de la empresa o del establecimiento, los totales no se pueden obtener por agregación de los datos detallados ya que las categorías de actividad o municipio desconocido no se presentan en las tablas.

Conceptos y definiciones estadísticos

A. Unidad Jurídica

Las unidades jurídicas son personas jurídicas cuya existencia está reconocida por la ley independientemente de las personas o instituciones que las posean o que sean miembros de ellas, o personas físicas que, en calidad de independientes, ejercen una actividad económica. La unidad jurídica sigue constituyendo, sola o a veces junto con otras unidades, el soporte jurídico de la unidad estadística "empresa".

B. Empresa

Es la combinación más pequeña de unidades jurídicas que constituye una unidad organizativa de producción de bienes y servicios y que disfruta de una cierta autonomía de decisión, principalmente a la hora de emplear los recursos corrientes de que dispone. Ejerce una o más actividades en uno o varios lugares. Una empresa puede corresponder a una única unidad jurídica.

En determinadas circunstancias puede corresponder a la reunión de diversas unidades jurídicas, de manera que alguna de ellas ejercen actividades exclusivamente en beneficio de otra entidad legal y su existencia sólo se explica por razones administrativas sin que por ello sean significativas desde el punto de vista económico. Pertenece asimismo a esta categoría una gran parte de las unidades jurídicas sin empleo. Con frecuencia se trata de actividades auxiliares de las actividades de la unidad jurídica matriz a la que secundan, a la que pertenecen y a la que deben vincularse para constituir la entidad "empresa", utilizada para el análisis económico.

C. Unidad local

Corresponde a una empresa o a una parte de empresas (taller, fábrica, almacén, oficinas, mina, depósito, etc.) sita en un lugar delimitado topográficamente. En dicho lugar o a partir de él se realizan actividades económicas a las que -salvo excepciones- dedican su trabajo una o varias personas (llegado el caso, en jornada parcial) por cuenta de una misma empresa.

En el caso en que una persona trabaje en varios lugares o trabaje en el domicilio, la unidad local de las que depende es el lugar desde el que recibe las instrucciones o en el que se organiza el trabajo. Ha de poderse precisar el empleo que está adscrito a toda unidad local. No obstante, toda unidad jurídica -desde el momento en que sirve de apoyo jurídico a una empresa o a una parte de empresa- debe contar con una unidad local sede, aunque no trabaje nadie allí. Por otra parte, una unidad local puede agrupar exclusivamente actividades auxiliares.

Un lugar delimitado topográficamente debe entenderse en sentido estricto: dos unidades de una misma empresa que tienen localizaciones diferentes (incluso en el seno de la circunscripción administrativa más pequeña del Estado miembro) deben considerarse como dos unidades locales. Sin embargo, puede ocurrir que la misma unidad local pueda encontrarse topográficamente en varias circunscripciones administrativas contiguas. En ese caso, se ha convenido que sea determinante la dirección postal.

D. Forma jurídica de la empresa

La forma jurídica de la empresa se obtiene a partir del primer carácter del número de identificación fiscal (NIF). En este caso se excluyen las comunidades de propietarios en régimen de propiedad horizontal.

E. Actividad económica

La actividad económica viene codificada por la CNAE 2009. En esta operación se excluyen las actividades 84 (Administración Pública y defensa; Seguridad Social obligatoria), 94 (Actividades asociativas), 97 (Actividades de los hogares como empleadores de personal doméstico), 98 (Actividades de los hogares como productores de bienes y servicios para uso propio) y 99 (Actividades de organizaciones y organismos extraterritoriales).

Ámbito territorial

El ámbito territorial utilizado es la Comunidad de Cantabria, recogiéndose aquellas unidades locales que ejerzan alguna actividad en el territorio de la misma

Ámbito temporal

La información de partida con la que se construye el Directorio de Empresas de Cantabria procedente de fuentes administrativas tiene como fecha de referencia el 31de Diciembre de cada año.

Variables de clasificación

Las variables utilizadas que permiten obtener desgloses y clasificaciones de las distintas poblaciones estadísticas son las siguientes:

- Forma jurídica de la empresa.
- Estrato de empleo de la unidad local y de la empresa.
- Actividad económica principal de la unidad local y de la empresa.
- Unidades geográficas en las que se ubica la unidad local: Municipios.

Fuentes de información

El directorio de empresas de Cantabria se actualiza a partir de la información recabada de diversas fuentes:

- Registro de centros de cotización de la Seguridad Social
- Registro de declarantes del Impuesto de Actividad Económica
- Directorio de Centros Educativos de la Consejería de Educación de Cantabria
- Directorio de Centros de Atención Primaria de Cantabria
- Encuestas propias.

El hecho de partir de unas informaciones que ya figuran en unos determinados registros administrativos informatizados que nos informan de los procesos de creación/desaparición de ciertas unidades o hechos económicos no supone por sí mismo, por mera trascripción o fotocopia, la elaboración de un repertorio "estadístico" de unidades locales de actividad económica. Un directorio requiere, como mínimo, una identificación territorial precisa del lugar en el que efectivamente se realiza la actividad, una adecuada codificación de la actividad económica y una serie de indicadores de tamaño, requisitos que no siempre se encuentran presentes en los registros de partida y que casi nunca se mantienen suficientemente actualizados.

Para conseguir una mayor precisión en la localización del empleo de los establecimientos de Cantabria se ha utilizado la información obtenida de la encuesta realizada desde el ICANE para el año 2.009 a 2.012 a 800 empresas, a partir de una muestra estratificada por actividad y empleo.

Intercambio de datos

Existe un Convenio de colaboración entre el Instituto Nacional de Estadística (INE) y el ICANE para la adopción de protocolos de intercambio de directorios estadísticos que regula la transmisión y el uso de la información requerida, con el fin de contrastar y mejorar la consistencia entre los directorios de ambos organismos (BOC 116, del viernes 15 de junio de 2012).

Procedimiento de actualización

La lógica de actualización del Directorio depende de la dimensión temporal de los datos, así como de una serie de interrelaciones entre entidades (personas físicas o entidades jurídicas) y fuentes de información. Asimismo, la ejecución de determinados procesos solamente es posible si la información se encuentra homogeneizada y validada por otros procesos. De todo ello resulta un flujo de ejecución que debe respetarse, a pesar de que cada módulo software es independiente entre sí como código ejecutable.

En primer lugar, hay que tener en cuenta la dimensión temporal. Si se dispone de datos correspondientes a varios períodos, es preciso realizar la actualización anual siguiendo el orden cronológico. De otro modo, fallarán algunas reglas de selección/eliminación basadas en la presencia/ausencia en determinadas fuentes en períodos distintos.

A continuación se describen los pasos seguidos en la elaboración del Directorio.

A) Carga de fuentes

La información procedente de los ficheros fuente se almacena en una estructura común, *origen*, sobre la cual se aplican la mayor parte de los procesos previos a la incorporación de los nuevos datos al Directorio.

El tratamiento automatizado de información procedente de diversas fuentes, requiere de procedimientos de unificación de formatos de datos y codificaciones, con el fin de posibilitar las comparaciones entre las características de interés pertenecientes a una misma entidad. El conjunto de tratamientos destinados a este fin se conoce como *normalización*.

Entre los procesos de normalización cabe destacar los siguientes:

- Asignación de identificadores únicos a cada registro, y de metadatos que permitan su localización en las fuentes
- Unificación de formatos de datos, codificación de caracteres, capitalización y segmentación de direcciones postales
- Corrección y depuración de información ausente, errónea o incoherente entre fuentes

B) Aplicación de reglas de selección/eliminación

Las entidades que pasarán a formar parte del directorio han de cumplir una serie de requisitos, establecidos en función del registro administrativo en que se encuentren, ramas de actividad y empleo remunerado. Los registros que no satisfacen dichos requisitos se marcan como eliminados, permaneciendo en *origen* para verificaciones posteriores.

C) Procesado de direcciones postales

Este paso es clave para la correcta asignación de las unidades locales (*establecimientos* en adelante). Dado que no existe un identificador para los establecimientos, como el CIF en el caso de las empresas, es necesario utilizar la dirección postal para determinar la identidad de los registros de cada empresa en diferentes fuentes.

La comparación automática de las direcciones postales supone un formidable problema, puesto que la información suele ser incompleta, imprecisa y presentar diferentes formatos entre las fuentes. Los métodos basados en la igualdad entre los distintos *elementos* que conforman una dirección postal (calle, número, piso, código postal, etc.) no pueden aplicarse, debido a la

heterogeneidad de los datos, siendo necesario construir *modelos probabilísticos* y aplicar técnicas de *aprendizaje de máquina* para realizar la identificación.

El proceso de comparación de las direcciones se realiza en dos etapas. En la primera, conocida como *segmentación* o *normalización*, se extraen los elementos de las direcciones originales para incorporarlos a una estructura única, basada en la especificación europea INSPIRE. Las direcciones quedan así convertidas en *vectores*. En la segunda etapa, *deduplicación*, se aplica una *métrica* a cada par de vectores de dirección, en función de cuyo valor se establece si ambos corresponden a la misma dirección ó no.

Normalización
El proceso de normalización se lleva a cabo empleando la herramienta febrl, desarrollada por el Data Mining Group de la Universidad Nacional de Australia (ANU, por sus siglas en inglés). La normalización se lleva a cabo mediante el modelado de las direcciones usando un *modelo oculto de Markov*. En este tipo de modelos, partimos de suponer que cada dirección es una secuencia de estados (tipo de vía -> nombre de vía -> portal), pero, al leerlas, desconocemos en qué estado concreto nos encontramos, conociendo únicamente el conjunto de observaciones.

Previo a cada ejecución del software de normalización, se recrea el modelo de Markov a partir de un *conjunto de entrenamiento* que contiene una muestra de direcciones correctamente normalizadas. El modelo resultante es una matriz de *probabilidades* de estados y transiciones, condicionadas por las secuencias de estados extraídas del conjunto de entrenamiento.

Deduplicación
Para la búsqueda de duplicados, lo primero es subdividir en *bloques* el conjunto total de direcciones, de manera que sólo hagamos comparaciones entre direcciones del mismo bloque. Los bloques se definen por identificador de la empresa (NIF/CIF), código de provincia y código de municipio.

A continuación, se realizarán una serie de comparaciones entre los elementos postales, usando funciones de similitud (por ejemplo, que los códigos postales sean iguales, que en el portal no haya más de cinco números de diferencia, o que la distancia Jaro-Winkler sea mayor al 75%). A cada comprobación se le asignará una puntuación máxima (concordancia completa), una puntuación mínima (discordancia completa) y una puntuación missing (en caso de que uno o los dos valores estén en blanco).

Una vez computada la puntuación de similitud de las dos direcciones, sumando las puntuaciones obtenidas en las diferentes comparaciones de campos, se compara este número con un valor umbral: en caso de superarse ese umbral, se determinará que las direcciones son lo suficientemente similares como para ser duplicados; en caso contrario, se dirá que son diferentes.

El último paso consiste en evaluar la calidad de cada dirección (que dependerá del número de campos con valor y la longitud de estos) para determinar, de forma efectiva, qué dirección se considera *duplicada* y cuál *original*.

En cada una de las etapas del proceso, se verifica la integridad y coherencia de la información mediante una serie de pruebas de validación con una doble finalidad: por un lado, asegurar la correcta aplicación de la "lógica de negocio" y, por otro, detectar errores o excepciones en los datos fuente que pueden depurarse antes de pasar a la siguiente etapa.

Simbología

(..) El dato no existe o no esta disponible
(*) Secreto estadístico