1.4.- Tratamiento estadístico.

El tratamiento estadístico y la explotación de datos del Registro Mercantil se inscribe dentro de un procedimiento general diseñado por el Instituto de Estadística de la Comunidad de Madrid, que se denomina PROTEO, y que determina la metodología aplicable a este tipo de operaciones.

- **Preparación de ficheros**. Recoge varios aspectos centrados en la transformación de los ficheros originales y homogeneización de los mismos en formato único, el cruce de los distintos ficheros muestrales, la consolidación de las unidades monetarias en las que vienen expresados los datos originales, la generación y tratamiento del código de actividad, la eliminación de registros de sectores no investigados, y, en el caso de los registros procedentes de los ficheros del Registro Mercantil, la detección de registros duplicados o que contienen información consolidada, el control de la estructura formal y el cruce con directorios. Dentro de esta primera fase se considera también un primer proceso de estimación o generación de variables en situaciones obvias (sumatorios sin cumplimentar y errores evidentes de signo) y un proceso de estimación inicial del empleo ante la no existencia del dato.

Esta primera fase consiste básicamente en la obtención de un fichero muestral apto para abordar las fases de tratamiento posteriores.

- **Depuración y estimación del fichero muestral de empresas.** Es una fase compleja consistente en procesos secuenciales para la detección y estimación de datos inconsistentes que finaliza con la obtención de un fichero muestral de empresas completo y válido. Básicamente la secuencia se puede resumir en: detección de valores inconsistentes (valores erróneos o ausencia de dato) generando un sistema completo de calificadores para la totalidad de variables del fichero, estimación de datos calificados como 'pendientes de estimar' y procesos de ajuste.

La articulación de los procesos es esencial en esta fase y contiene un conjunto de subfases cuyos elementos básicos se pueden resumir en:

1. Subfases de detección determinística de errores. Basados en la utilización de controles apriorísticos de coherencia lógico-aritmética que instrumentalmente generan los campos de calificación de cada variable. Dentro de estas subfases se incluye la validación determinística básica sobre campos obligatorios, signos no admitidos, etc.

2. Subfases de estimación determinística de variables. Se trata de estimaciones simples y centradas en casos muy concretos de campos con calificador 'pendiente de estimar' a través de información externa.

3. Subfases de detección y corrección no probabilística de errores. La herramienta utilizada se denomina QUANTITA y ha sido diseñada para la detección de valores inconsistentes a partir de un conjunto de controles apriorísticos que son definidos previamente y su estimación posterior a través de un proceso de imputación general basado en los mismos criterios. En el caso concreto de ésta operación, la definición

de controles se realiza a través del cálculo inicial de correlaciones, y la posterior obtención de límites. La herramienta sigue el criterio del cambio mínimo y permite un uso interactivo y/o automático. Esta utilidad facilita el tratamiento individualizado de los registros más significativos de cara a los resultados finales.

4. Subfases de detección y/o corrección probabilística de errores. Se utiliza la herramienta denominada MEGA (Método de Elevación General Asistida). Básicamente se trata de una herramienta que permite la estimación máximo-verosímil individual de cada variable a partir de un conjunto de factores de estimación definidos a partir del estudio de correlaciones y que incluye la parametrización de la cobertura mínima, de las variables de estratificación y del método de cálculo del estimador. La combinación de ambos métodos de detección de inconsistencias (probabilística y no probabilística) permiten aseguran la coherencia horizontal y vertical del fichero.

5. Subfases de ajuste. Permite la reestimación de valores de forma determinística de acuerdo a un esquema algebraico simple de tipo lineal. La estimación es proporcional con restricciones y permite asegurar la coherencia aritmética de cada registro derivada de la propia estructura de las cuentas.

- **Regionalización del fichero muestral de empresas y elevación del colectivo.** El último bloque de procesos está encaminado a la obtención del fichero completo de unidades locales ubicadas en la región conteniendo, para cada una de ellas, información completa de balance y de la cuenta de pérdidas y ganancias (modelo abreviado). Para ello es necesario, en primer lugar, la obtención de un fichero muestral de establecimientos y, en segundo lugar, la extrapolación al colectivo total.

La transformación de la muestra de empresas a muestra de establecimientos requiere inicialmente un proceso de cruce con el Directorio de Actividades Económicas de la Comunidad de Madrid. En este directorio se dispone de información completa sobre la identificación de las empresas a las que pertenecen los establecimientos, localización de los mismos, actividad principal que desarrollan las unidades locales, tipología de los establecimientos (sede central, establecimiento productivo, sede y establecimiento productivo, o establecimiento auxiliar), y empleo. El casamiento, a excepción de los registros investigados directamente en campo por el Instituto, básicamente se realiza a través del CIF y exige tareas de validación que derivan en investigaciones específicas.

Posteriormente se distribuyen los datos económicos de la empresa entre los establecimientos localizados en la región, en la mayoría de los casos vía empleo ya que es la única variable disponible para la información proveniente del registro mercantil, en este primer trabajo. En posteriores ediciones se podrá utilizar igualmente toda la batería de variables del balance abreviado y la cuenta de pérdidas y ganancias abreviada para este cometido.

El tercer bloque de procesos aborda la elevación del colectivo con la filosofía de que la extrapolación de la muestra al colectivo no es más que un proceso de estimación de registros en los que se dispone de un mínimo de información (básicamente empleo y

actividad), a partir de otros sobre los que se dispone de información completa (muestra válida).

Bajo estas premisas los procesos son similares a los desarrollados en la depuración y estimación de la muestra de empresas. Los registros muestrales de establecimientos heredan los calificadores del fichero muestral de empresas, calificadores que intervienen en los procesos de elevación realizados con MEGA. De forma recurrente se realizan subfases de detección de inconsistencias y ajustes similares a las anteriormente descritas.

El último paso consiste en el cálculo y estimación del colectivo final. La existencia de esta fase responde a la imposibilidad de disponer de un Directorio exhaustivo y completamente actualizado que no exija la estimación del colectivo total a partir de información externa.

Como resultado final de esta fase es la obtención de un fichero completo y consistente del colectivo sobre el que se realiza la tabulación.

Además de la tabulación diseñada, el proceso seguido permite un abanico muy amplio de tabulaciones adicionales, a mayor detalle o con cruces no considerados, que se puede solicitar por los cauces ordinarios.