

Diseño de una muestra representativa de la población andaluza para encuestas presenciales basada en el catastro

Manuel Trujillo Carmona
mtrujillo@iesacsic.es
IESA/CSIC

Resumen

El diseño de una muestra representativa de la población de un territorio requiere previamente disponer de un marco o listado para elegir de forma aleatoria de este marco a los individuos que formarán parte de la muestra.

Centrándonos en encuestas presenciales, así lo hacen los institutos oficiales de estadística, como el INE o el IECA, y en algunos casos otras encuestas como la Encuesta Social Europea o algunas de las que realiza el CIS. En todos estos casos el marco utilizado es el Padrón de Habitantes, que gestiona el INE. Sin embargo, al ser información confidencial, la política del INE o el IECA para ceder los datos del INE es muy restrictiva, y no es posible utilizar esta fuente para la realización de la mayoría de encuestas.

La alternativa habitual es entonces la utilización de métodos que intenten buscar una aleatoriedad en el recorrido, sin tener un marco explícito, por ejemplo con un sistema de rutas aleatorias dentro de las secciones censales, como hace normalmente el CIS o ha venido haciendo también el IESA. El problema de estos sistemas es que es difícil de controlar a priori las direcciones exactas que teóricamente entran en la muestra, con la dificultad por tanto de controlar el trabajo de campo que hace el encuestador, hacer revisitas, y la imposibilidad también de enviar anuncios de que se va a realizar la encuesta con antelación. Esto nos lleva por tanto a que no se pueda garantizar con estos métodos una encuesta probabilística, hablándose de encuestas semi-probabilísticas.

En este artículo se muestra la base de datos del catastro como una alternativa de marco muestral accesible a cualquiera, prácticamente exhaustivo y sin problemas de confidencialidad, que puede servir para realizar encuestas probabilísticas a la población de forma presencial. También se mostrará cómo se ha utilizado por parte del IESA para la captación de los encuestadores para el panel de encuestas que está preparando, Panel Ciudadano PACIS.

Además del catastro se utilizan también como fuentes auxiliares la explotación estadística del

Padrón de Habitantes para secciones y núcleos de población, los resultados del Censo de Población de 2.011 para secciones, el Callejero Digital Unificado de Andalucía del IECA, el callejero del INE, la delimitación de núcleos de población del IECA, la delimitación geográfica de secciones del INE correspondiente al Censo de 2.011, y la delimitación geográfica de códigos postales, obtenida de Cartociudad.

El trabajo de campo se está llevando a cabo en el último trimestre de 2.014 y por tanto no podemos dar resultados sobre cómo está funcionando.

Palabras clave: muestreo, catastro, muestras representativas, callejeros

Introducción

El diseño de una muestra representativa de la población de un territorio requiere previamente disponer de un marco o listado para elegir de forma aleatoria de este marco a los individuos que formarán parte de la muestra.

Centrándonos en encuestas presenciales, así lo hacen los institutos oficiales de estadística, como el INE o el IECA, y en algunos casos otras encuestas como la Encuesta Social Europea o algunas de las que realiza el CIS. En todos estos casos el marco utilizado es el Padrón de Habitantes, que gestiona el INE. Sin embargo, al ser información confidencial, la política del INE o el IECA para ceder los datos del INE es muy restrictiva, y no es posible utilizar esta fuente para la realización de la mayoría de encuestas.

La alternativa habitual es entonces la utilización de métodos que intenten buscar una aleatoriedad en el recorrido, sin tener un marco explícito, por ejemplo con un sistema de rutas aleatorias dentro de las secciones censales, como hace normalmente el CIS (Rodríguez Osuna, 1991) o ha venido haciendo también el IESA. El problema de estos sistemas es que es difícil de controlar a priori las direcciones exactas que teóricamente entran en la muestra, con la dificultad por tanto de controlar el trabajo de campo que hace el encuestador, hacer revisitas, y la imposibilidad también de enviar anuncios de que se va a realizar la encuesta con antelación. Esto nos lleva por tanto a que no se pueda garantizar con estos métodos una encuesta probabilísticas, hablándose de encuestas semi-probabilísticas.

En este artículo se muestra la base de datos del catastro como una alternativa de marco muestral accesible a cualquiera, prácticamente exhaustivo y sin problemas de confidencialidad, que puede servir para realizar encuestas probabilísticas a la población de forma presencial.

También se mostrará cómo se ha utilizado por parte del IESA para la captación de los encuestadores para el panel de encuestas que está preparando, Panel Ciudadano PACIS.

Además del catastro se utilizan también como fuentes auxiliares la explotación estadística del Padrón de Habitantes para secciones y núcleos de población, los resultados del Censo de Población de 2.011 para secciones, el Callejero Digital Unificado de Andalucía del IECA, el callejero del INE, la delimitación de núcleos de población del IECA, la delimitación geográfica de secciones del INE correspondiente al Censo de 2.011, y la delimitación geográfica de códigos postales, obtenida de Cartociudad.

Objetivos de la encuesta y diseño muestral

Objetivos de la encuesta

La muestra cuyo diseño se describe tiene como objetivo ofrecer una selección probabilística de personas para el Panel Ciudadano de Andalucía, marco de personas que servirá de base para futuras encuestas del IESA, y cuyos datos se recogerán mediante Internet y telefónicamente, y dependiendo del porcentaje de personas que no puedan responder a las encuestas mediante estas dos vías, también podría tener una parte de recogida presencial.

La captación se lleva a cabo en el cuarto trimestre de 2.014.

Para conseguir la mayor tasa de respuesta posible, y que la muestra sea lo más aleatoria posible, nos planteamos prescindir del sistema que habíamos utilizado hasta ahora, de rutas aleatorias y cuotas, e ir a una selección puramente probabilística.

En primer lugar tratamos de realizar la selección basándonos en el Padrón de Habitantes, pero tanto por parte del INE como por parte del IECA se nos contestó que solo podrían hacernos muestras con este registro si se tratara de estadísticas oficiales o dentro de convenios internacionales, que no era el caso. Por tanto finalmente optamos por la novedosa estrategia de utilizar el catastro como marco.

Dado que puede haber encuestas referidas a todo el hogar, y no a personas, vamos a considerar nuestro universo los hogares andaluces, incluyendo todos aquellos que residan de hecho en Andalucía en el momento de la captación, independientemente de su residencia legal. Intentaremos incluir dentro del panel a todas las personas mayores de edad que formen parte de los hogares seleccionados.

A nivel de error, el objetivo es tener un marco suficiente para realizar las encuestas futuras del IESA, que se plantean en general con unas necesidades del nivel de error que podrían llegar al 2%, al nivel de confianza del 95%, para toda Andalucía.

Procedimiento de captación

Se diseñará una página web para la presentación y gestión del panel en el que los ciudadanos invitados a participar podrán incorporar su información de contacto a partir de un código personalizado que se le dará con la invitación. El proceso de captación / invitación se hará en las siguientes fases:

1ª Fase: Envío postal de tríptico informativo del panel y carta de invitación para participar en el mismo. En dicha carta se ofertarán distintas vías para darse de alta en el mismo, acceso online y llamada a línea 900.

2ª Fase: Recuerdo postal a las direcciones válidas que no hayan contactado con el IESA (cartas que no hayan sido devueltas pero que tampoco hayan llegado a contactar).

3ª Fase: Captación personalizada. Encuestadores se encargarán de contactar con los hogares incluidos en la muestra que no hayan contactado con el IESA. Los encuestadores llevarán las direcciones de cada vivienda seleccionada en la muestra, y un plano con su localización.

Tipo de muestreo

El muestreo va a ser bietápico, con una primera selección de secciones censales (y en algunos casos del núcleo de población), y una segunda selección de viviendas dentro de las secciones censales. El número de viviendas elegido en cada sección será a partir de 20, y está en función de la dificultad esperada de conseguir respuestas de los habitantes de la sección, como se explicará más adelante.

Además, las secciones se estratificarán por provincia y según su comportamiento político en las elecciones autonómicas de 2.011, con una clasificación realizada con análisis de conglomerados que las divide en 7 grupos.

El seccionado utilizado es el de 2.011, ya que para algunos procesos nos interesaba tener los datos del censo, y queríamos tener una misma fecha de referencia para poblaciones y

callejeros.

Tamaño muestral

El tamaño muestral que vamos a buscar sea aquel que nos permita posteriormente obtener muestras para encuestas con un error típico máximo del 3%, lo que da suponiendo muestreo aleatorio simple, una muestra de 1.056 entrevistas. Por experiencias anteriores, el muestreo por conglomerados utilizando secciones en Andalucía acarrea un efecto del diseño usualmente hasta de 1,6 (Trujillo Carmona, 2.001). Por tanto, habría que multiplicar la cifra anterior por este número, quedando en 1.690 el tamaño muestral buscado. Partiendo de esta cifra, llegamos a un tamaño de direcciones necesario de unas 10.000 viviendas, mediante el siguiente proceso, explicado a la inversa:

De 10.000 direcciones de viviendas extraídas del marco, según el censo de 2.011, aproximadamente un 15% aproximadamente están vacías, y un 15% son viviendas secundarias, lo que sumaría un 30% de viviendas donde podría no haber nadie. Como en verdad algunas de las secundarias no serán tales, sino que pueden ser viviendas donde resida alguien del hogar, vamos a estimar que es un 20% el porcentaje de direcciones elegido donde no hay nadie. Por tanto nos quedamos con 8.000 viviendas habitadas en la muestra.

A las 10.000 viviendas elegidas se les incluye en el proceso de captación, en la fase 1, y se les envía una carta de presentación. Las cartas devueltas pasarán un proceso de revisión de su dirección, viendo si la vivienda realmente existe, y si hay una dirección alternativa.

Posteriormente se les envía una carta de recuerdo, excepto a aquellas que hayan contestado. Como ya hemos dicho, de las 10.000 iniciales, estimamos que son 8.000 aquellas viviendas donde habita alguien que puede leer estas cartas. De estas 8.000 calculamos, por los antecedentes conocidos, que una pequeña parte se apuntará al panel sin necesidad de más recuerdos (Struminskaya y otros, 2014). Esta pequeña parte la ciframos en un 5%. Por tanto, nos quedan 7.600 hogares que pasarán a la 3ª fase.

De esta cifra, la proporción esperada de personas que acepten participar en el panel estaría entre el 40% y el 50%, por lo que sumando los 400 hogares reclutados anteriormente nos daría un tamaño de muestra de hogares de entre 3.440 y 4.200 (tasa de participación de entre el 43% y el 53%). En cualquier caso el objetivo mínimo para dar validez al proceso sería una tasa de participación del 30% (2.400 hogares). De estos hogares se seleccionaría posteriormente los requeridos para las muestras.

Selección de las secciones

En principio el planteamiento era realizar la selección de las secciones con probabilidad proporcional al número de viviendas principales que nos da el censo de viviendas de 2.011. Sin embargo, tras una primera inspección se puso de relieve que había bastantes secciones donde no había una estimación del número de viviendas principales, y en algunos casos ni siquiera del número total de viviendas. Y en muchos casos donde sí había estimación, ésta era poco verosímil. Tras pedir explicaciones al INE, se nos confirma que estas variables proceden del muestreo, y por tanto pueden estar en algunos casos bastante alejadas de la realidad. Como alternativa, utilizamos una selección proporcional al número de personas en cada sección que nos da el Padrón de Habitantes de 2.011.

Selección del núcleo de población

Hay secciones que tienen varios núcleos de población, algunos de ellos de características muy distintas entre sí, por ejemplo la proporción de viviendas secundarias. Para hacer más fácil el trabajo de campo, y para controlar las características de la muestra, es preferible, cuando sea posible, elegir solo un núcleo en cada sección para muestrear, o si no fuera posible, más de uno.

Para seleccionar el núcleo se ha empleado la población residente en el cruce de sección y núcleo procedente del Padrón de Habitantes de 2001, que se pidió al IECA. Una vez asignadas las parcelas a cada núcleo de población o diseminado, como se explicará más adelante, se seleccionó dentro de cada sección un núcleo. Esta asignación en algunos casos no ha sido del todo aleatoria, porque se encontraron algunos núcleos de población con una proporción de viviendas vacías o secundarias muy altas, que hubiera hecho casi imposible encontrar una muestra allí. Por tanto se han desechado aquellos núcleos o diseminados en que el cociente entre viviendas y población fuera muy alto –en algunos casos se ha encontrado hasta 8 viviendas por persona-. También se han desechado aquellos núcleos con un número de habitantes o de viviendas inferior a 50, porque no podían suministrar suficiente muestra. De los núcleos de cada sección que no se desechan se ha realizado una selección aleatoria con probabilidad proporcional a su número de habitantes.

Asignación del número de cartas a enviar en cada sección

El objetivo debería ser que el tamaño muestral en cada sección sea el mismo, para así

conseguir que la probabilidad de entrar en la muestra sea la misma para cada hogar. Sin embargo, es difícil controlar a priori el número de cartas que tenemos que enviar, ya que no sabemos qué porcentaje de viviendas de las localizadas en el catastro van a estar habitadas, ni tampoco sabemos el porcentaje de colaboración que obtendremos en cada una de las secciones.

Con criterios de suficiente dispersión, pero no excesiva, se han seleccionado 400 secciones, con lo cual el número medio de cartas por sección será de 25. Según los objetivos, de entre el 43% y el 53%, de tasa de respuesta, tendríamos entre 10 y 13 hogares por sección. Según el mínimo establecido para dar por válido el proceso, de 2.400 hogares en total, tendríamos que tener un mínimo de 6 hogares en cada sección.

Para asignar un número de cartas a cada sección hemos tenido en cuenta únicamente el primero de los factores, y el reparto ha sido así: en primer lugar, se asigna un número mínimo de 20 cartas por sección. Este número sería el adecuado si todo funcionara bien, siendo necesaria una tasa de respuesta del 30% para llegar al mínimo de 6 hogares. Como se han seleccionado 400 secciones, esto suman 8.000 direcciones. Las 2.000 restantes se reparten, dependiendo del dato de que se disponga, en función de la proporción entre viviendas principales y viviendas totales, para las secciones en las que se disponga de este dato para 2.011, o de la proporción entre viviendas encontradas en el catastro y población, y con un coeficiente multiplicativo de 2 o 1,5 para los casos en que fuera más compleja la ubicación de las viviendas. Sobre todo, para viviendas en parcelaciones donde no están muy claras las direcciones, o en urbanizaciones residenciales donde hay abundancia de extranjeros y viviendas de segunda residencia y calles tortuosas. El objetivo es enviar más cartas a aquellas secciones donde con más probabilidad va a haber más cartas que no lleguen y más rechazos.

En caso de que en alguna sección no se llegara al mínimo de 6 hogares en el panel se seleccionarían viviendas sustitutas.

Descripción de la base de datos del catastro

El catastro mantiene una base de datos donde figuran todo tipo de bienes inmobiliarios. La consecuencia es que es una base de datos muy compleja. Además, existe por un lado la versión alfanumérica, y por otro lado la versión cartográfica.

Conseguir la base de datos de cualquier municipio es muy sencillo, basta con ir a la página web de la dirección electrónica del catastro <https://www.sedecatastro.gob.es> e ir al servicio de

descarga masiva de datos (es necesario certificado electrónico), donde tenemos las opciones de descarga de información alfanumérica (formato CAT) y descarga de cartografía vectorial (formato Shapefile). El inconveniente es que la descarga es municipio a municipio. Además, está dividido en catastro de urbano y catastro de rústica, tanto en un caso como en otro.

Para evitar la tediosa tarea de descargar cada municipio, hicimos una petición a la dirección de contacto de catastro que amablemente, y en poco tiempo, nos enviaron ambas bases de datos para toda Andalucía. Eso sí, dividida en distintos ficheros para cada municipio y separadas las bases de datos de rústica de la de urbana, tanto para las bases de datos alfanuméricas como las cartográficas.

Paso ahora a describir someramente la estructura y el formato de cada base de datos:

Base de datos alfanumérica

La información se nos remitió dividida por municipios, y cada uno de ellos divididos en rústica y urbana.

La base de datos alfanumérica consta de 5 tablas, aunque están todas condensadas en un solo fichero Excel, que se trata con 5 plantillas de Excel que importan el fichero de forma adecuada a cada tabla, identificada con una variable llamada tipo de registro. Una vez importado, hay que eliminar la información correspondiente al resto de tablas, eliminando los tipos de registro restantes.

Las tablas son:

Tipo 11: Registro de Finca. Existirá uno por cada parcela catastral implicada.

Tipo 13: Registro de Unidad Constructiva. Existirá uno por cada unidad constructiva en cada parcela catastral.

Tipo 14: Registro de Construcción. Existirá uno por cada construcción de cada unidad constructiva en cada parcela catastral

Tipo 15: Registro de Inmueble. Existirá uno por cada bien inmueble en cada parcela catastral

Tipo 17: Registro de cultivos

Con un municipio concreto se importaron todas las tablas, tanto para el catastro de rústica como para el de urbana, y se analizó cuales podrían servir para nuestros objetivos, llegando a la conclusión de que la tabla 15, Registro de Inmueble, contenía toda la información necesaria para la identificación de las viviendas.

Esta tabla contiene numerosas variables. Entre ellas, las que nos van a servir son:

Identificadores

- Parcela catastral
- Número identificador del bien dentro de la parcela

Localización del inmueble

- Código de provincia
- Código de municipio
- Nombre de municipio
- Nombre de la entidad menor
- Código de vía pública
- Tipo de vía pública
- Nombre de la vía pública
- Primer número de policía (número del portal)
- Primera letra (para duplicados)
- Kilómetro
- Bloque
- Escalera
- Planta
- Puerta
- Textos complementarios de dirección
- Código postal
- Nombre del paraje

Datos de filtro de los inmuebles

- Tipo de bien (rústico, urbano, especial)
- Uso de los bienes inmuebles

Tratamiento

Como ya se ha dicho, las tablas vienen separadas en distintos archivos por cada municipio, y en archivos .cat de texto, que se deben importar con una plantilla de Excel. Para no hacerlo uno por uno, se unieron todos los archivos de cada provincia en uno solo, mediante el comando de Ms-DOS Copy. Estos archivos se abrieron con la mencionada plantilla, y se eliminaron los campos que no correspondían con el tipo de archivo 15. Se guardan estos archivos como Excel y se importan a Access.

En la base de datos del catastro aparecen todo tipo de parcelas. Sin embargo, a nosotros solo nos van a interesar aquellas destinadas a viviendas, que se deben corresponder con el código

de uso "V". Ello no nos garantiza que estén habitadas, y mucho menos que sean viviendas principales, pero sí al menos que, para catastro, están destinadas a este fin. Por tanto, eliminamos todos los registros que tienen otros usos.

Base de datos cartográfica

La gran ventaja del catastro es que además de tener la dirección postal de la vivienda, también podemos localizarla perfectamente, ya que se proporcionan capas geográficas en formato .shp. De nuevo tenemos un paquete cartográfico para cada municipio, y además separado entre rústica y urbana.

En cada paquete tenemos numerosas capas, entre las que podemos elegir las que queremos representar. Las principales son:

Capas del parcelario catastral:

- Masa: agrupación de parcelas (manzanas de urbana y polígonos de rústica)
- Parcela: parcelas catastrales
- Subparcela: Subparcelas de cultivos (zonas con los mismos cultivos)
- Constru: Subparcelas urbanas que representan los volúmenes edificadas en una parcela.

Capas con información geográfica auxiliar

- Carvia: Nombre de las vías (tabla de dbf, no geográfica)
- Ejes: Ejes de elementos lineales (calles, carreteras,...)
- Elemlin: Elementos cartográficos lineales
- Elempun: Elementos cartográficos puntuales
- Elementex: Rótulos del mapa
- Limites: Límites administrativos

Aquí la tabla fundamental es la tabla Parcela, ya que es en la que se podremos reflejar las viviendas enlazándola con la base de datos alfanumérica. La tabla constru es interesante también en la fase de depuración para ver si las viviendas seleccionadas están realmente construidas.

VARIABLES PRINCIPALES DE LA TABLA PARCELA:

- Mapa: Mapa en el que se encuentra la parcela
- Municipio: Código de municipio
- Masa: Referencia de la manzana o polígono

- Tipo: Urbana, rústica, diseminado o dominio público
- Parcela: Número de parcela dentro de la masa
- Refcat: referencia catastral de la parcela
- CoorX: Coordenada X del centroide de la parcela
- CoorY: Coordenada Y del centroide de la parcela
- Vía: Código de vía (se relaciona para obtener el nombre con la tabla CARVIA)
- Número: Número de portal
- Númerodup: Indicador para números duplicados
- Numsymbol: Número del símbolo de representación
- Area: Superficie de la parcela
- Fechaalta: Fecha de creación de la parcela en la base de datos
- Fechabaja: Fecha de borrado de la parcela en la base de datos
- PCAT1: Primeros siete números de Refcat
- PCAT2: Últimos siete números de Refcat

Tratamiento

El programa utilizado para tratar la información geográfica ha sido Q-GIS. En primer lugar se han colocado en una carpeta por provincia todos los archivos de la capa parcela de esa provincia. Se han unido formando una sola capa de tipo .shp por provincia mediante el comando “combinar archivos shape en uno” de Q-GIS.

Las tablas de Carvia, que nos servirán para obtener la dirección postal con este archivo, las unimos utilizando la utilidad para Excel ASAP (<http://www.asap-utilities.com/>), y las importamos a Access.

Asignación de la sección censal

La dispersión de la población, y el hecho de que la captación vaya a ser principalmente en los propios hogares de los encuestados hace inviable un muestreo aleatorio simple. Por tanto se va a utilizar como conglomerado, al igual que se hace en casi todos los procesos de muestreo en hogares, la sección censal como unidad primaria de muestreo. El problema que tenemos en nuestro caso es que las bases de datos de catastro no tienen asignada la sección censal a que pertenecen.

Una primera posibilidad sería utilizar la dirección postal para asignar la sección, intentando relacionarla con el callejero de secciones del INE. Sin embargo, tras un primer análisis se vio que las direcciones postales del catastro y las del INE difieren mucho, y por tanto iba a ser muy

complicado establecer esa relación.

Como ya se hablará más adelante, una segunda posibilidad era utilizar como nexo de unión el portalero del Callejero Digital Unificado de Andalucía, que teóricamente incluye tanto referencia catastral como la relación con el callejero del INE, lo que nos daría la sección censal. Esta opción no resulta válida para gran parte de las direcciones, y por tanto también se descarta.

Finalmente, la opción elegida ha sido utilizar la cartografía del seccionado censal, y adjudicar por procedimientos de asignación espacial cada parcela a la sección en que se encuentre el centroide de la parcela catastral. Ello puede llevar a algún error, primero por falta de exactitud de la cartografía, y también porque la dirección postal de la parcela puede estar en distinto lugar que la sección. Pero estos errores no serán un porcentaje muy alto, y pueden ser corregidos en el proceso de revisión. Se ha utilizado la cartografía de secciones censales del Censo de 2.011, que se puede descargar de la web del INE.

Para asignar los centroides se ha utilizado el algoritmo Polygon centroids, de la caja de herramientas SAGA, y para asignar la sección a cada parcela, la herramienta “unir atributos por localización”.

Una vez asignada la sección censal, hacemos la correspondencia en Q-GIS con la base de datos de parcelas, y ya nos quedamos en adelante solo con las parcelas correspondientes a las secciones seleccionadas, lo cual nos agiliza mucho el procesamiento.

Asignación del núcleo de población

La base de datos de catastro no incluye una asignación de núcleos de población, salvo que sean entidades menores, lo cual no es lo más habitual. Por tanto hay que buscar otra fuente. El Callejero Digital o el del INE tienen el mismo problema ya comentado para las secciones. La solución buscada es utilizar también la asignación espacial, pero en este caso nos encontramos con que el nomenclátor del INE, es decir, el reparto oficial de núcleos y entidades singulares, no tiene su correspondiente base cartográfica. Como solución, utilizamos la capa de núcleos del DERA-100 del Instituto de Estadística de Andalucía. El DERA-100 contiene dos capas de núcleos distintas, la su-01 y la su-03, siendo su-01 la original del IGN y su-03 una más detallada. Se utilizará esta, aún a sabiendas de que a veces lo que se marca como núcleo no son sino barrios. También quedarán parcelas sin asignar a ningún núcleo, que normalmente corresponderán con diseminados. La forma de asignación espacial será la misma que la del seccionado explicada anteriormente.

Una vez asignados estos núcleos se ha buscado la correspondencia con los núcleos del INE, de los cuales tenemos datos. Esto se ha hecho de forma manual, tanto utilizando los callejeros del INE como de forma visual y otras formas de encontrarlos, tarea que a veces no ha sido fácil.

Asignación del Código Postal

Dado que el primer contacto con la vivienda seleccionada va a ser postal, es fundamental que el código postal sea correcto. La base de datos de catastro incluye el código postal pero está vacío en gran parte de los casos. Para asignarlo, hemos utilizado una correspondencia entre localidades y códigos postales que nos ha enviado Correos, para las localidades que solo tienen un código postal. Para el resto, y para aquellos casos en que no se correspondían nuestras localidades con las de correos, de nuevo hemos utilizado la asignación espacial.

En principio se ha utilizado la capa de códigos postales que aparece en Cartociudad, y hemos procedido con la misma técnica utilizada anteriormente para secciones y núcleos. Sin embargo, al revisarlo, nos hemos dado cuenta de que había numerosas viviendas en que no coincidía esta asignación espacial con los resultados que se obtiene al buscar el código postal en la propia web de correos, concluyendo que esta capa, que es la misma que se obtiene por ejemplo en la descarga del DERA-100, es errónea o está anticuada. Buscando, en el servicio WMS del IECA del DEA-100 (Divisiones administrativas), existe otra capa distinta de código postal que sí se ajusta más a los resultados que da la web de correos. Sin embargo, esta capa no he encontrado la posibilidad de descargarla. No dudo que nos habría sido enviada por el IECA, pero como esta dificultad ha aparecido a última hora, por evitar demoras, se ha hecho la revisión de los códigos postales erróneos de forma manual, tanto utilizando la web de Correos como esta capa. Constatando además que también muchos de los códigos postales que aparecen en la base de datos de catastro son erróneos y que, no muchos, pero aún hay alguno erróneo en la capa del DEA-100.

Callejero Digital de Andalucía Unificado

Otra fuente que hemos utilizado para revisar las direcciones, comprobar los núcleos de población y como dirección alternativa es el Callejero Digital de Andalucía Unificado.

Este callejero está disponible para su descarga en la página web del IECA

<http://www.juntadeandalucia.es/institutodeestadisticaycartografia/callejero/utilidades.htm>

Se presenta con dos tablas, una de viales y otra de portales, aunque la de portales ya incluye toda la información que necesitamos. Por supuesto, la tabla de portales no incluye nada sobre el número de viviendas que hay en cada portal, por tanto no serviría como base para el muestreo. También hay que tener en cuenta que los portales no siempre tienen que ser viviendas, sino que pueden ser otro tipo de locales.

La estructura de la tabla de portales es la siguiente:

- Id_Vial: identificador propio del vial
- INE-Vía: Código del callejero del INE
- DGC-Vía: Código del callejero de Catastro
- TVIAN: Tipo de vía (nombre corto)
- NOM_TIP_VI: Tipo de vía (completo)
- NOM_VIA: Nombre de la vía
- Sobrenombre: Variantes del nombre
- ID_POR_PK: Identificador del portal
- TIPO_PORTA: Portal, punto kilométrico o diseminado
- NUM_POR_DESDE: Número de portal
- EXT_DESDE: Letra para duplicados
- NUM_POR_HASTA: Para los casos en que un portal abarca varios números.
- EXT_HASTA: Letra para duplicados en ese caso
- BLOQUE: Número de bloque (casos de conjuntos donde hay un solo número pero varias entradas)
- PORTAL: Número de portal (casos de conjuntos donde hay un solo número pero varias entradas)
- ESCALERA: Número de escalera (casos de conjuntos donde hay un solo número pero varias entradas)
- REFCATPARC: Referencia catastral de la parcela
- TXT_APP: Otros textos aclaratorios sobre la dirección
- NOM_TIPO_AGRUP: Urbanización, polígono,...
- NOM_AGRUP: Texto aclaratorio del lugar donde se encuentra el portal
- INE_NUCLEO: Código de núcleo INE
- NOM_NUCLEO: Nombre del núcleo en que se encuentra el portal
- INE_MUN: Código de municipio INE
- NOM_MUNICIP: Nombre del municipio INE
- COD_POSTAL: Código postal
- X: Coordenada X
- Y: Coordenada Y

Las coordenadas permiten importar la tabla a Q-GIS y representar geográficamente todos los portales.

Con esta estructura en teoría se podría ligar perfectamente tanto con el catastro a través de la referencia catastral de la parcela como con el callejero del INE a través del código de vía. Sin embargo, casi la mitad de los portales carece de referencia catastral, y a la inversa, hay aproximadamente un 50% de las parcelas catastrales cuya referencia catastral no aparece en el CDAU. Estudiándolo un poco, hemos visto que se trata sobre todo de aquellos portales donde no coinciden las direcciones, especialmente el nombre de las calles.

El campo de callejero INE solo está vacío en aproximadamente un 10% de los casos, por lo que aquí sí que sería más fácil la conexión.

Dado que el número de portales sin referencia catastral es tan alto se ha optado por asignarlo según su localización geográfica, en esta ocasión vamos a asignar el portal (capa puntual) a la parcela (capa poligonal) en que esté inserto, con la herramienta “unir atributos por localización” que hemos utilizado antes. Como en muchas ocasiones el portal está justo en el borde de la parcela, y podría no entrar en la parcela, lo hemos convertido en un “punto gordo” con la herramienta buffer, y así nos aseguramos que parte del punto entre en la parcela. De esta forma se han asignado un 94% de portales a parcelas.

Este callejero nos sirve para encontrar direcciones alternativas en los casos en que las cartas nos vengán devueltas, y también para ayudar al encuestador a encontrar la vivienda.

Cobertura y fiabilidad

El número de viviendas en la base de datos de catastro es de 4.263.721. Según el INE, en 2.001, el número de viviendas es de 4.350.095. Tendríamos por tanto una cobertura muy alta, superior al 95%. Aunque el número de viviendas del INE también podría ser inferior al real. Se ha encontrado que en determinados casos el número de viviendas en el catastro es muy inferior al esperado según la población empadronada. Sobre todo hemos encontrado que esto se da en dos circunstancias: viviendas irregulares, y cuevas. En ambos casos, parece que hay muchas personas empadronadas en viviendas ausentes del catastro.

También pueden darse casos de duplicidades, ya que en ocasiones un inmueble del catastro podría corresponderse con más de una vivienda, o una vivienda con más de un bien del catastro. Estos desajustes, que alterarían la probabilidad de inclusión de la vivienda, no

creemos que sean cuantitativamente importantes.

Otra posible fuente de sobrecobertura es las viviendas que aún no están construidas, o promociones que, estando construidas, no se han entregado. Estos desajustes llevarían a un alto número de devoluciones de cartas, o de tiempo perdido por los encuestadores, pero no alterarían la fiabilidad.

La duda principal es que la variable de uso se corresponda con la realidad. En este caso de nuevo seguramente será más habitual la sobrecobertura, por viviendas que realmente estén usándose como negocios o como oficina, que lo contrario, aunque también podría haber. El problema que más nos ha preocupado, como se ha explicado, ha sido el de las direcciones postales, porque son muchísimas las direcciones postales de catastro que no coinciden con las reales. Por ello se ha recurrido a otras fuentes de datos tanto para los nombres de las calles como para los códigos postales, que sirvan para utilizarse en caso de devolución de cartas.

Conclusión

El Catastro puede tomarse como marco para un muestreo probabilístico alternativo en caso de no poder utilizarse el Padrón de Habitantes. Sin embargo, su utilización es compleja tanto por la estructura de su base de datos como por la necesidad de utilizar técnicas especiales para enriquecerla con datos censales, padronales, etc...

El hecho de no conocer qué vivienda es usada como residencia principal constituye el principal inconveniente, que sin duda encarecerá el trabajo de campo, y la revisión de los nombres de calles, códigos postales, etc... requiere de la utilización de otros callejeros. También el desconocimiento del nombre de los residentes es un factor en contra de la participación en el panel. Pero frente a la alternativa de realización con rutas aleatorias, es mucho mejor ya que permite controlar exactamente el lugar donde debe realizarse la captación (entrevista, si se tratara de una encuesta), permite control de la entrevista, revisitas, etc...

Referencias

Rodríguez Osuna, J., *Métodos de muestreo*. CIS, Madrid (1.991)

Struminskaya, B., L. Kaczmirek, I. Schaurer, y W. Bandilla., *Assesing representativeness of a probability-based online panel in Germany*. En Callegaro, M, R. Balkar, J. Bethlehem, A. Goritz, *Online panel research: a data quality perspective*. Wiley, Chichester, UK (2.014)

Trujillo Carmona, M., *Optimización del error muestral en encuestas a población general: criterios de estratificación y cálculo del error con 'wesvar complex samples'*, Metodología de Encuestas vol. 2 n° 2 (2.001).