DISEÑO MUESTRAL PARA UNA ENCUESTA TELEFÓNICA A NIVEL EUROPEO

Comunicación presentada al III Congreso de Metodología de Encuestas
Granada, 15 al 17 de septiembre de 2004

Oscar Antonio Santacreu Fernández
Dpto. Sociología II, Psic., Comunic. y Didáctica
Universidad de Alicante
Apartado Correos 99
E – 03080 Alicante
e-mail: oscar.santacreu@ua.es

Abstract

En el contexto del Proyecto Europeo Pioneur se ha planteado la necesidad de aplicar una encuesta telefónica en cada uno de los cinco países participantes en el proyecto (Italia, España, Francia, Reino Unido, Alemania). Para ello, resultaba indispensable disponer de los datos telefónicos del subconjunto de la población objetivo en cada uno de los citados países. Habida cuenta de que dicha información no se puede obtener de forma directa, se ha planteado un mecanismo que ha permitido la elaboración de dicha información.

La selección de la muestra se ha realizado tomando como elemento identificativo de la nacionalidad el nombre y el apellido de los individuos, mediante varios pasos diferenciados. En primer lugar se han obtenido, a partir de repertorios telefónicos en soporte DVD, las frecuencias de las unidades lingüísticas (nombre y apellidos) de cada país tomando en consideración la distribución geográfica de las unidades lingüísticas en base al código postal como elemento adicional de control, eliminando aquellas zonas que pudieran ocasionar problemas (Alsacia-Lorena, por ejemplo). A partir de dichas frecuencias se ha calculado el índice de probabilidad de cada una de esas unidades lingüísticas respecto al hecho de pertenecer a un determinado país. Tras una segunda validación, se ha programado un algoritmo para la detección y extracción de las distintas muestras nacionales desde los repertorios telefónicos en DVD. A continuación se han desarrollado y aplicado unos filtros para depurar la muestra obtenida, y finalmente, otro algoritmo ha separado el número de teléfono de la información identificativa (nombre, dirección...) y ha mezclado aleatoriamente dichos números, a fin de obtener unos archivos que garanticen el anonimato de los entrevistados.

Contexto de la investigación

La presente comunicación muestra el procedimiento aplicado para la resolución de un problema planteado en el contexto del proyecto de investigación europeo PIONEUR (Pioneers of Europe's Integration "from below": Mobility and the emergence of European Identity among National and Foreign Citizens in the EU), financiado por la Comisión Europea dentro del V Programa Marco.

Dicho proyecto plantea que un creciente número de políticas de la Unión Europea se basan en asumir que un incremento en la movilidad de las personas entre los estados miembros conllevará un aumento de la productividad, reducirá el desempleo, hará que las competencias profesionales sean más accesibles y acelerará el proceso de integración europea. Sin embargo, la movilidad continúa siendo extremadamente baja: en la actualidad, sólo el 1,5% de los ciudadanos europeos viven en un país distinto de su país de origen. Además, la mayoría de ellos proceden de las migraciones laborales de la posguerra y tienen un perfil distinto al que tienen emigrantes más recientes.

El proyecto investiga 1) las condiciones y motivaciones de la movilidad interna en la Unión Europea, 2) los efectos de esta movilidad interna en las expectativas y en la calidad de vida de los protagonistas de esta movilidad, 3) el impacto de la movilidad interna y externa en las actitudes hacia la Unión Europea y en la identificación con sus instituciones.

En términos operativos, se distinguen tres grupos de residentes europeos como objeto de análisis:

- Naturales: ciudadanos europeos que residen en el estado miembro del cual son ciudadanos;
- Móviles internos: ciudadanos de la UE que residen en un estado miembro distinto del suyo propio;
- Móviles externos: ciudadanos de países no pertenecientes a la UE, de países del Centro y Este de Europa, residentes en un estado miembro.

Inicialmente, cada uno de estos grupos se investiga por separado. Los naturales a través del análisis secundario de datos; los móviles internos a través de un estudio etnográfico cualitativo y una encuesta cuantitativa; los móviles externos a través de un análisis de contenido de grupos para, a continuación, proceder a un análisis comparativo inter-grupo (en particular, entre naturales y móviles internos, y entre móviles internos y externos). Estos análisis tienen como meta conocer el impacto que tienen los distintos tipos de movilidad en la construcción de la identidad europea y del bienestar de las personas. El estudio busca, además, identificar las características individuales y contextuales necesarias como precondición para la movilidad interna en la Unión Europea.

La red de trabajo está constituida por los siguientes grupos de trabajo:

- Observatorio Europeo de Tendencias Sociales (OBETS) – Universitad de Alicante – España
- Centro Interuniversitario di Sociologia Politica (CIUSPO) – Universidad de Florencia – Italia
- Centre for Socio-Legal Studies (CSLS) – Universidad de Oxford – Reino Unido
- Centre d'Etude de la Vie Politique Française (CEVIPOF) – CNRS – France
- Zentrum für Umfragen, Methoden und Analysen (ZUMA) – Mannheim – Germany

Como se ha dicho, uno de las técnicas a aplicar en el proyecto consiste en la elaboración de una encuesta, suponiendo la mayor fuente de datos del proyecto: 5.000 encuestas a ciudadanos de la Unión Europea que estén viviendo en otro país de la UE. Concretamente, la encuesta se lleva a cabo en Italia, España, Reino Unido, Francia y Alemania, e incluye cuatro nacionalidades por país, esto es,

ciudadanos de Alemania, Reino Unido, Francia, Italia y España, excluyendo, naturalmente, la nacionalidad del país en el que se realiza la encuesta.

Las tareas implicadas en este paquete de trabajo son:

1. Organización de una reunión de trabajo con expertos para la elaboración de la base metodológica.
2. Diseño del borrador del cuestionario para las entrevistas estructuradas, y traducción a los cinco idiomas de los países en los que se realizará la encuesta.
3. Creación de las muestras
4. Pretest del cuestionario
5. Selección de empresas para el campo, siendo requisito el entrenamiento de entrevistadores bilingües.
6. Entrevista de 1.000 individuos (250 casos por nacionalidad) en cada país
7. Tabulación y combinación de los datos obtenidos en cada país.
8. Análisis

Para el punto correspondiente a la creación de las muestras se pensó inicialmente en conseguir listados de los consulados en cada país, pero esta opción resultó ser inviable. Cabía, por tanto, idear un mecanismo que permitiera la obtención de dicha información. Dado que el medio elegido para la realización de las entrevistas fue el sistema CATI (entrevistas telefónicas), la solución mas obvia fue acudir a las guías telefónicas, aunque esto planteó un problema adicional: la necesidad de diseñar un procedimiento que permitiera extraer la información relativa a los extranjeros de las nacionalidades objeto de estudio. Se presenta aquí el procedimiento seguido para la obtención de la muestra.

Procedimiento

A partir de los datos disponibles en los listados telefónicos, el único criterio viable para discriminar la nacionalidad de los sujetos es la utilización del nombre y el apellido o apellidos de los individuos como elementos identificativos del país de origen, lo que nos lleva a otra necesidad previa: un criterio que nos permita asignar a un determinado nombre o apellido una probabilidad de pertenecer a una determinada nacionalidad. Los pasos seguidos para conseguirlo fueron los siguientes:

1. Obtención de los datos telefónicos en bruto correspondientes a Francia, Alemania, Italia, Reino Unido y España en un formato que permita su tratamiento estadístico.
2. Análisis de frecuencias de unidades lingüísticas por país
3. Cálculo de probabilidades.
4. Extracción de la muestra.
5. Revisión y testado de los datos.

Así, el primer paso consistió en adquirir una guía telefónica que incluyera los cinco países que nos ocupan –Italia, Alemania, Francia, Reino Unido y España– y que permitiera la exportación de sus registros en formato texto (ASCII),con el único fin de permitir su tratamiento estadístico mediante un procedimiento informático automatizado.

La información típica recogida en las guías telefónicas constan de los siguientes campos:

- Nombre y apellido(s)
- Dirección
- Código postal
- Población
- Número de teléfono

Para la extracción se consideraron únicamente los registros correspondientes a particulares, y se desecharon aquellos correspondientes a empresas. Igualmente, se desecharon los campos de Dirección y Población (excepto en el caso inglés, donde se desechó el código postal y se mantuvo el de población por razones que se explican más adelante). De este modo, los registros disponibles para cada país fueron los que se muestran en la siguiente tabla:

Tabla 1. Registros telefónicos disponibles por país

País	**Registros**
Francia	20.473.368
Italia	16.941.708
España	12.411.976
Reino Unido	11.451.423
Alemania	30.590.871

Fuente: elaboración propia

A continuación se realizó un análisis estadístico de los datos exportados a texto. Así se determinó que el nombre aparecía en el primer campo de cada registro, delimitado por el carácter tabulación (ASCII 9), y que los elementos dentro de cada campo estaban separados por el carácter espacio (ASCII 32). Nombres y apellidos aparecían separados por un carácter visualmente idéntico al espacio, pero distinguible para los algoritmos (el carácter ASCII 160) lo que permitía distinguir entre nombre y apellido (o apellidos). En el caso español, la configuración más frecuente resultó ser aquella en la que el campo nombre contenía dos apellidos y la inicial del nombre seguido por un punto. En el caso francés, la mayor parte de los registros incluían en el campo nombre el tratamiento (MR, MME...) seguido del apellido y del nombre. Muchos registros correspondían a matrimonios e incluían el apellido de cada uno de los cónyuges precedidos del tratamiento, sin incluir el nombre. En el caso alemán, el registro nombre contenía mayoritariamente un apellido y un nombre, si bien en algunos casos aparecían dos individuos en el mismo registro. En el caso italiano, la configuración más frecuente resultó ser la de un apellido seguido de uno o de dos nombres. En el caso inglés, la configuración más frecuente correspondía a un apellido seguido del nombre y, en algunos casos, de un segundo nombre abreviado.

A la vista de estas particularidades, se desarrolló un sistema de algoritmos de filtrado *ad hoc* mediante el cual se extrajeron, de cada registro, dos unidades lingüísticas correspondientes, según el país, a dos apellidos o a un apellido y un nombre. Estas unidades lingüísticas se ordenaron en una lista por país a partir de la cual se calcularon las frecuencias.

Como elemento adicional de control, se tomó en consideración la distribución geográfica de los apellidos y nombres dentro de cada país. Para ello se utilizó como variable de agrupación las dos primeras cifras del código postal a fin de segmentar las listas, de modo que los resultados fueron organizados por grupos correspondientes a zonas postales, obteniendo así un listado de los apellidos más frecuentes por zona y favoreciendo de este modo la heterogeneidad de la distribución de la muestra a nivel geográfico. En el caso inglés, la especial configuración de los códigos postales hizo necesario realizar la segmentación geográfica a partir de los nombres de las poblaciones.

Esta segmentación geográfica permitió también excluir aquellas zonas que podían producir problemas en la identificación de los apellidos y nombres nacionales, como por ejemplo Alsacia-Lorena o las zonas germanófonas del norte de Italia.

Una vez programado el sistema, compuesto de varios módulos, se recorrió cada uno de los listados realizando las tareas que a continuación se relatan, proporcionando, para cada país y cada zona

geográfica (z), un conjunto de resultados compuesto de las unidades lingüísticas correspondientes a nombres y apellidos:

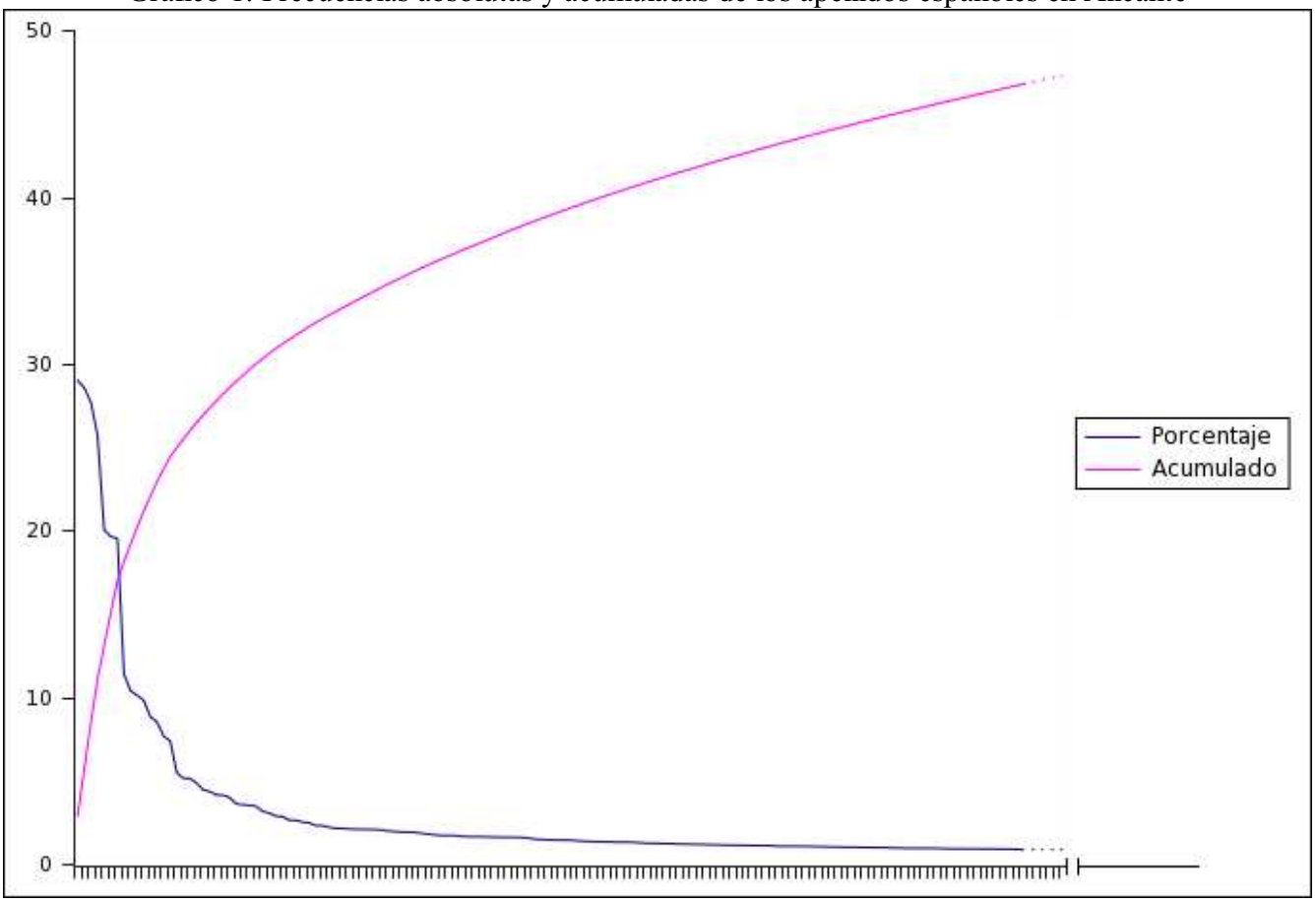
$$\Omega_z=\{unidad1,unidad2,unidad3...unidad\mathit{n}\}$$

Nuestro objetivo era conseguir el subconjunto de unidades lingüísticas "nacionales" con más probabilidades de pertenecer a cada zona geográfica (z) dentro de un país:

$$A_z=\{unidad1,unidad2...\}$$

Ordenando los apellidos de mayor a menor frecuencia, observamos que los apellidos más frecuentes suponían la mayor parte de las frecuencias, por lo que se tomó como elementos del subconjunto A_z aquellas unidades lingüísticas que estaban por encima de un determinado percentil en la zona geográfica **z**. En el Gráfico 1 se observa el caso de Alicante.

Gráfico 1. Frecuencias absolutas y acumuladas de los apellidos españoles en Alicante



Fuente: elaboración propia

En el gráfico, la línea descendente se refiere al porcentaje que presenta un apellido, del más al menos frecuente. La línea ascendente representa el porcentaje acumulado para ese apellido. Se observa por tanto que son los apellidos más frecuentes los que, naturalmente, incluyen a un mayor porcentaje de la población, disminuyendo la aportación en términos de población para los apellidos conforme nos alejamos hacia la cola de la distribución.

Por otro lado, el subconjunto A_p de cada país resulta de la unión de los subconjuntos correspondientes a cada una de las zonas de dicho país:

$$A_p=\{A_z^1 \cup A_z^2 \cup A_z^3 \cup ... \cup A_z^n\}$$

Seguidamente, el algoritmo calculó la probabilidad para cada uno de los elementos del subconjunto A_p partiendo del concepto de frecuencia relativa $f_r=k/n$ siendo n el número de unidades lingüísticas y k las apariciones de la unidad lingüística (regla de Laplace):

$$P(A)=(\text{casos favorables})/(\text{casos posibles})$$

Se seleccionaron entonces aquellas unidades lingüísticas que tenían una mayor probabilidad y se enviaron a los distintos equipos de la red de investigación para que cada uno revisara de forma manual el listado de nombres correspondientes a su nacionalidad a fin de eliminar posibles errores.

Llegados a este punto habíamos conseguido una valiosa información: qué unidades lingüísticas (nombres y apellidos) identifican, y con qué probabilidad, la nacionalidad de un individuo cualquiera de la guía telefónica, de modo que podíamos pasar a la siguiente etapa: la extracción de la muestra. En este proceso participaron los siguientes elementos:

- A_o es el subconjunto A_p del país de origen.
- A_d es el subconjunto A_p del país de destino.
- A_c es el subconjunto A_p de todos los países excluyendo el país de destino
- Ω_{pd} es el conjunto de la población en el país de destino

Para realizar dicho proceso, el algoritmo eliminó, en primer lugar, los elementos de A_o que aparecen en A_c, es decir:

$$A_o = A_o - (A_o \cap A_c)$$

Esto quiere decir que si nos encontramos, por ejemplo, con el nombre *María* como indicador de nacionalidad española, dicho nombre sería extraído de A_o dado que se encuentra también, por ejemplo, en el subconjunto de Alemania o de Reino Unido.

Realizada esta tarea previa, el algoritmo recorrió el listín telefónico de cada país, registro a registro, realizando las siguientes tareas:

a) Identificar los distintos campos que componen dicho registro, guardando cada uno de ellos en una variable.
b) Eliminar los componentes referidos al tratamiento (por ejemplo, MME., MR., DR.,etc)
c) Estandarizar los caracteres contenidos en las variables (eliminar acentos, diéresis, etcétera de modo que los únicos caracteres presentes sean A-Z). Esto es necesario para evitar que "González" y "Gonzalez" sean consideradas palabras distintas por el acento de la primera.
d) Para acelerar el proceso (idea surgida a partir de los pretest), contar el número de apellidos, como factor determinante para la identificación de españoles en otros países, así como de extranjeros en España.
e) Comprobar que el primer apellido se encuentra en A_o, es decir, si corresponde al país de origen, y en caso contrario saltar al siguiente registro. De esta forma se acelera el proceso al no seguir analizando de forma innecesaria el resto de los elementos del nombre.
f) Detectar la presencia de más elementos del subconjunto A_o, de modo que a cada registro del subconjunto Ω_{pd} se le asocia una probabilidad de pertenecer al país de origen en función de

la suma de las probabilidades de los elementos del subconjunto A_o presentes en dicho registro.
Por ejemplo, Antonio González tiene más probabilidades de ser español que John González porque se suman las probabilidades de la unidad lingüística *González* y de la unidad lingüística *Antonio.* O dicho en otras palabras, si en un registro del subconjunto Ω_{pd} se encuentran dos elementos A_o^1 y A_o^2, a dicho registro se le asignaría la siguiente probabilidad.

$$P(A_o^1 \cup A_o^2) = P(A_o^1) + P(A_o^2) - P(A_o^1 \cap A_o^2)$$

g) El programa compara a continuación las unidades lingüísticas con las correspondientes a los otros países (A_c) y resta, en su caso, dicha probabilidad para penalizar los nombres y apellidos que pudieran pertenecer a otra nacionalidad:

$$P(A_o^1) = P(A_o^1) * (1 - P(A_c))$$

h) Si la probabilidad del registro es mayor que una probabilidad determinada para cada nacionalidad en función de varios pretest, el programa guarda la siguiente información del registro en un nuevo archivo temporal:
   a. El número de coincidencias encontradas
   b. La probabilidad de pertenecer a la nacionalidad objetivo
   c. El nombre
   d. El código postal
   e. El número de teléfono

A continuación, el programa elimina los registros correspondientes a las regiones antes mencionadas (Alsacia-Lorena y norte de Italia, por ejemplo) para evitar falsos positivos en la realización de las entrevistas.

El programa extrae entonces un listado de los nombres y apellidos contenidos en el archivo de muestra, para su revisión visual. Este procedimiento demostró ser extraordinariamente eficaz para garantizar la calidad de la muestra.

Por último, el programa toma el archivo temporal de muestra una vez revisado, lo ordena de mayor a menor utilizando los campos *coincidencias* y *probabilidad* y extrae dos archivos, que serán los definitivos.

- **Archivo 1**: Contiene cuatro campos: coincidencias, probabilidad, número de teléfono y código postal. Los registros del archivo se desordenan aleatoriamente antes de su utilización en el sistema CATI
- **Archivo 2**: Contiene el campo probabilidad truncado a 1 decimal y el campo nombre. Los registros del archivo se desordenan aleatoriamente a fin de evitar la identificación con el correspondiente registro del *archivo 1*, garantizando de ese modo el anonimato del entrevistado.

El proceso proporcionó para cada unos de los grupos de investigación listados de entre 5.000 y 10.000 contactos telefónicos para cada nacionalidad objetivo de estudio, que demostraron ser suficientes para las 250 entrevistas necesarias por nacionalidad, en todos los casos excepto en el de Reino Unido, que se comenta a continuación.

Problemas y soluciones

Tuvimos un problema importante en el caso de Reino Unido, el único país desde el que nos comentaron que los listados proporcionados producían demasiados falsos positivos. Tras varias revisiones tuvimos que llegar a la conclusión de que el problema no residía en el método de extracción, sino en la calidad de los repertorios telefónicos de aquel país y en las particulares características de su mercado de telecomunicaciones. La única solución posible fue la optimización de los listados (menos números pero más probables), y en mayor medida la búsqueda de contactos mediante la técnica del *snow ball* o bola de nieve, es decir, pedir a cada uno de los contactados que proporcionara, si estaba en su mano y si lo tenía a bien, nuevos contactos dentro del perfil de la encuesta, que a su vez proporcionaran nuevos contactos.

Otro de los problemas a los que nos tuvimos que enfrentar fue la lentitud del algoritmo, dado que tenía que comparar, en muchos casos, miles de nombres para cada uno de los registros. La solución vino a partir de numerosas revisiones y optimizaciones del código (escrito en PHP Script), como por ejemplo tener en cuenta el número de apellidos, o generar listas temporales por iniciales de modo que cada nombre era comparado únicamente con los nombres con la misma inicial, o otras optimizaciones cuyo contenido sobrepasa los objetivos de la presente comunicación. En todo caso debo resaltar que resultó ser un trabajo laborioso pero a la vez altamente satisfactorio por los resultados obtenidos.