8. Nuevas perspectivas en la explotación y aprovechamiento de los datos secundarios

Benjamín González Rodríguez

1. Preliminares

1. Un nuevo contexto histórico-cultural para los datos secundarios

Desde que se publicara, hace poco más de cinco años, este capítulo del libro *El análisis de la realidad social*, han acaecido muchas cosas en el ámbito de la investigación social y, en concreto, con respecto al manejo de la información existente, genéricamente denominada *datos secundarios*. El acontecimiento más llamativo y, posiblemente, el más importante, se refiere a la aparición y generalización relativa del uso de Internet, a nivel internacional (no tanto en el caso de España), lo que ha supuesto una extensión significativa del acceso mucho más amplio, fácil y asequible a grandes bases de datos de todo el mundo y a numerosas fuentes de información que anteriormente eran inexistentes o inaccesibles. Es cierto que en Internet existe información abundante y, con frecuencia, abrumadora, pero también resulta mucha de ella colosalmente irrelevante. La rapidez vertiginosa del flujo de información queda patente si se observa que si uno hace una búsqueda determinada al día de hoy, con unos descriptores, obtiene una información. Si al día siguiente realiza la misma búsqueda, con idénticos descriptores, obtendrá seguramente resultados distintos. ¿Qué puede significar esto? Por una parte, una cierta inestabilidad temporal de la información y, por la otra, una acumulación progresiva y gigantesca de la misma. El desarrollo acelerado de los nuevos sistemas de difusión y manejo de la información conlleva, en general, considerables peli-

gros o problemas ante la avalancha de información a la que se puede acceder con sólo pulsar unas cuantas teclas u oprimiendo el puntero del *ratón*. El primero de estos peligros es el de perderse, sobre todo si no se va con una idea clara de qué es lo que se busca y se necesita, ya que, tras pocos minutos de *navegación pinchando* en unos cuantos *enlaces* podemos encontrarnos con toneladas de información en el disco duro del ordenador. Bien es cierto que en pocos años ha habido un intenso desarrollo de *buscadores* capaces de remitirse unos a otros *(metabuscadores)*. Sin ellos, el cúmulo de información de Internet resultaría completamente inútil. El siguiente problema tiene mucho que ver con la tarea de discernir entre lo relevante y lo accesorio, entre la información fidedigna y la equívoca, entre lo útil y lo inútil. De hecho, existen diversas publicaciones encaminadas a guiar sobre los criterios a tener en cuenta para calibrar estos aspectos específicamente en el caso de Internet, como la de Alexander y Tate (1999) encaminada a enseñarle al lector como evaluar y crear información de calidad en la red. Están apareciendo asimismo muchos libros sobre cómo utilizar Internet con fines de investigación.

2. Nuevo papel de los datos secundarios

El objetivo principal de este capítulo va encaminado a destacar la importancia creciente que, desde el punto de vista de la investigación social, han adquirido los datos secundarios, tanto desde una perspectiva teórica como aplicada e instrumental. El argumento principal que anteriormente se esgrimía a favor de la utilización de datos secundarios se basaba en que la generación de datos/información primaria sale más costosa en términos económicos, de tiempo, de organización y gestión, por lo que cada vez le iba a resultar más difícil al investigador individual producir sus propios datos y, en consecuencia, tendría que recurrir a datos proporcionados por otras fuentes. En definitiva, se encontraría el investigador ante un mal menor. No obstante, no se trata de un simple mal menor ni de cuestiones de mero coste económico, sino que nos hallamos ante un cambio o salto cualitativo, como se decía antiguamente. La mejora rápida de la plétora de información, su accesibilidad y abaratamiento brinda una oportunidad magnífica al investigador social para que pueda poner en juego su propia imaginación y creatividad investigadora, que puede verse potenciada con la disponibilidad de técnicas analíticas más potentes y amigables. Resulta, por ello mismo, desafortunado el nombre de datos secundarios, debido a sus connotaciones negativas: complementarios, supletorios, accesorios. El uso normal del término lo contrapone a principal, importante. En realidad, cuando hablamos de investigación secundaria nos referimos a datos existentes y disponibles, sin implicar ningún tipo de aspecto peyorativo. Al contrario, debido sobre todo a los avances y desarrollos de bases de datos, este tipo de investigación está adquiriendo cada vez mayor importancia y alcance, tanto teórico como metodológico, siendo cada día mayor el número de

investigadores que se insertan en esta corriente de investigación, formando grupos de investigación radicados en el mismo o en diferentes países. Bien es cierto que esta última tendencia no es tan frecuente entre los investigadores españoles, debido posiblemente a las dificultades de comunicación derivadas del idioma y al relativo (sub)desarrollo del uso de las nuevas tecnologías.

El tema del coste se ha trasladado a otros niveles de la producción de datos y se sitúa en el ámbito del diseño, creación, mantenimiento y gestión de las grandes bases de datos. En la actualidad, ha dejado de tener sentido (al menos en cuanto a coste económico y social) generar una cantidad formidable de investigaciones (datos) que se utilizan una única vez para producir un informe y/o una publicación, pasando los datos posteriormente, en el mejor de los casos, a formar parte de un banco de datos inutilizado. Si estas bases de datos no se utilizan, terminan convirtiéndose en inmensas cárcavas inservibles. La situación parece ir cambiando. Las consultas de estas bases de datos van creciendo notablemente. Es más, las instituciones españolas empiezan a darse cuenta de que es necesario rentabilizar (científica y socialmente) esa información. Un ejemplo de ello sería la convocatoria que el Centro de Investigaciones Sociológicas viene realizando estos últimos años, encaminada a la explotación de los estudios existentes en su banco de datos. Es un comienzo, aunque estemos lejos todavía de la situación en que el recurrir a las bases de datos existentes constituya una práctica habitual, suponiendo, claro está, que estas fueran más numerosas y asequibles en términos de disponibilidad y coste.

3. Necesidad de nuevos enfoques metodológicos y técnicos

La investigación social ha de enfrentarse con urgencia al impacto que sobre los modos de hacer investigación social van a ejercer (están ejerciendo ya) las nuevas vías y posibilidades que brinda la tecnología disponible al respecto. Los métodos y las técnicas deberían reciclarse con premura para adaptarse a los cambios sociales que se están produciendo. Los propios diseños de investigación y las estrategias metodológicas para abordar el mundo social deben ser revisadas a la luz de las nuevas realidades. Es evidente que, si lo vemos desde la óptica de los datos secundarios, el problema no es hoy el de utilizar y analizar una encuesta realizada hace años (su fiabilidad, validez, adecuación a nuestro objeto de investigación). El problema reside, vaya por caso, en cómo manejar simultáneamente y para un mismo proyecto múltiples encuestas que pueden existir en diferentes países, así como otras fuentes de datos que pueden incluir imágenes digitalizadas o incluso textos derivados de la prensa o transcripciones de varios grupos de discusión realizados posiblemente en contextos culturales muy distintos. Existe ya la posibilidad de enlazar esta información dispersa en un único y mismo proyecto de investigación. Aparte del análisis de la calidad de *datos* provenientes de fuentes tan dispares, se plantean, entonces, problemas técnicos como el de *conjugar* de una manera compacta toda esta información. En realidad, nos enfrentamos a una

necesidad de reforzar enfoques conceptuales, metodológicos y técnicos distintos de los clásicos.

4. Datos secundarios, teoría y creatividad científica

Simplificando mucho, podría decirse, siguiendo a Mochmann y Guchteneire, que *la utilización de viejos datos para nuevas ideas sería la descripción coloquial de lo que se ha denominado en términos técnicos análisis secundario* (p. 2). Lo que equivale a no convertir los almacenes de información en inmensas cárcavas inútiles. En este sentido, indican estos autores que la misma información primaria (la misma pregunta de una encuesta, por ejemplo) puede ser utilizada por otro investigador para fines bastante diferentes. Ponen un ejemplo. La siguiente pregunta de una encuesta, *¿Con qué frecuencia suele hablar usted de política con sus vecinos?*, puede utilizarla un investigador en la investigación primaria para construir un índice de interés por la política, mientras que otro investigador puede querer utilizarla para un fin completamente diferente: construir un índice de integración en el barrio. Los límites, en este sentido, se ubican sólo en la capacidad creativa e imaginativa del analista de datos secundarios. El problema está en que no existen guías para determinar cuáles son las preguntas nuevas y originales que podemos hacerle a los viejos datos, lo que nos conduce de inmediato a la necesidad de recurrir a la teoría como orientadora de nuestras búsquedas de información para dar respuesta a problemas sociales relevantes. El analista de datos secundarios debe plantear el marco teórico del estudio y las hipótesis de investigación. El papel de la teoría es decisivo si se quiere que los datos tengan sentido y significado. No hay que olvidar que los datos no hablan por sí mismos. Si a esto añadimos un buen análisis exploratorio podríamos propiciar mejor el *insight* que provocara indagaciones conspicuas sobre las diferentes realidades sociales de interés. Ahora bien, este proceso requiere nuevos mecanismos que permitan un análisis más interactivo de los datos. De hecho, empiezan a estar disponibles programas informáticos que permiten una representación visual de los mismos, de forma que el investigador pueda *simular* situaciones distintas del tipo de *¿qué pasaría si....?* Finalmente, parece claro que para llevar a cabo una investigación basada en datos secundarios, es necesario que el investigador posea un buen conocimiento y familiarización con su área de investigación. En todo caso, no hay que olvidar que en la investigación juega un papel importante, no siempre reconocido explícitamente, la *casualidad*. Cuando nos encontramos con algunos resultados de investigación no previstos y que no pueden interpretarse en función de las teorías en vigor y que pueden dar lugar a nuevas teorías, estamos ante un fenómeno que Merton designó allá en los sesenta *pauta de serendipidad*, que no debe confundirse con el simple azar, puesto que el *jugar* con los datos, darles vuelta, ponerlos en contradicción consigo mismos constituyen mecanismos propiciadores de resulta-

dos o perspectivas nuevas. Cuando el investigador está familiarizado con las fuentes de información existentes en su área, posee unos recursos valiosos para apreciar lagunas, carencias y debilidades de la misma, con lo que está en disposición de sugerir ideas para posibles proyectos de investigación (propios y ajenos), y a partir de ahí la contribución al avance científico puede ser mayor al ir rellenando esas lagunas y vacíos de investigación.

5. Importancia del análisis de los objetivos

Como en cualquier otro tipo de investigación, también aquí resulta imprescindible la claridad en cuanto a los objetivos que se persiguen, ya que si estos no están claros se corren graves riesgos: *confundir o equivocar el objetivo propicia la utilización de métodos inadecuados y puede llevar a pronósticos e inferencias desdichadas* (Glymour *et al.* 1997, p. 17). Estos autores tratan de ejemplificar *lo fácil que resulta dar respuestas acertadas a preguntas equivocadas (ib.)*, e insisten en que no hay reglas definitivas para llevar a cabo una buena definición de objetivos sino que depende del buen criterio del investigador. De nuevo es necesario contar con la ayuda de un buen marco teórico.

6. Progreso científico (teórico, metodológico y técnico)

El desarrollo y la utilización de grandes bases de datos ha supuesto un avance importante en el desarrollo de las ciencias sociales. Se da ahora la posibilidad de que los teóricos utilicen datos existentes para incorporarlos en el desarrollo de sus teorías, potenciando así la relación entre teoría e investigación[1]. Desde otro punto de vista, dice Lane (1990, p. 190) que *el valor de la información o de los centros de documentación depende siempre del contexto teórico de que se trate o del problema e hipótesis que se esté planteando. Sólo los avances de la teoría científica social pueden guiar el desarrollo de los bases de datos nacionales*. Por otra parte, el generar y almacenar la información de forma que pueda ser utilizada con facilidad y provecho por distintos investigadores ha supuesto retos técnico/metodológicos importantes, dando origen a estrategias de investigación nuevas y eficaces (meta-análisis, síntesis de investigaciones, minería de datos *[data mining]*, etc.). La línea de progreso científico iniciada a partir de la utilización de datos secundarios no ha hecho más que empezar y está sujeta a un dinamismo intenso, creciente y prometedor.

7. Datos secundarios e investigación comparativa

La existencia de extensas bases de datos facilita grandemente el análisis comparativo que, hasta hace poco no resultaba muy factible. Con los datos adecuados pueden llevarse a cabo comparaciones tanto a niveles locales, nacio-

nales e internacionales como en distintos períodos de tiempo. Ciertamente, toda ciencia es comparativa, si bien aquí entendemos la investigación comparativa en un sentido más restringido, como el intento de estudiar los rasgos y patrones comunes y diferenciales de distintas sociedades, así como los factores que los producen. Cada vez se tiene más en cuenta el contexto en el que se producen los fenómenos sociales, por lo que *las comparaciones transnacionales han servido cada vez con más frecuencia como un medio de lograr una mejor comprensión de distintas sociedades, de sus estructuras y sus instituciones. El desarrollo de este (tercer) enfoque ha coincidido con el crecimiento de la colaboración internacional interdisciplinar y de las redes de las ciencias sociales que se han visto potenciadas desde los años setenta mediante una serie de iniciativas a lo largo de toda Europa* (Hantrais 1996). La Unión Europea trata de potenciar proyectos que impliquen a varios países miembros con el fin de conocer los resultados de políticas distintas y ver si son o no transferibles de unos países a otros.

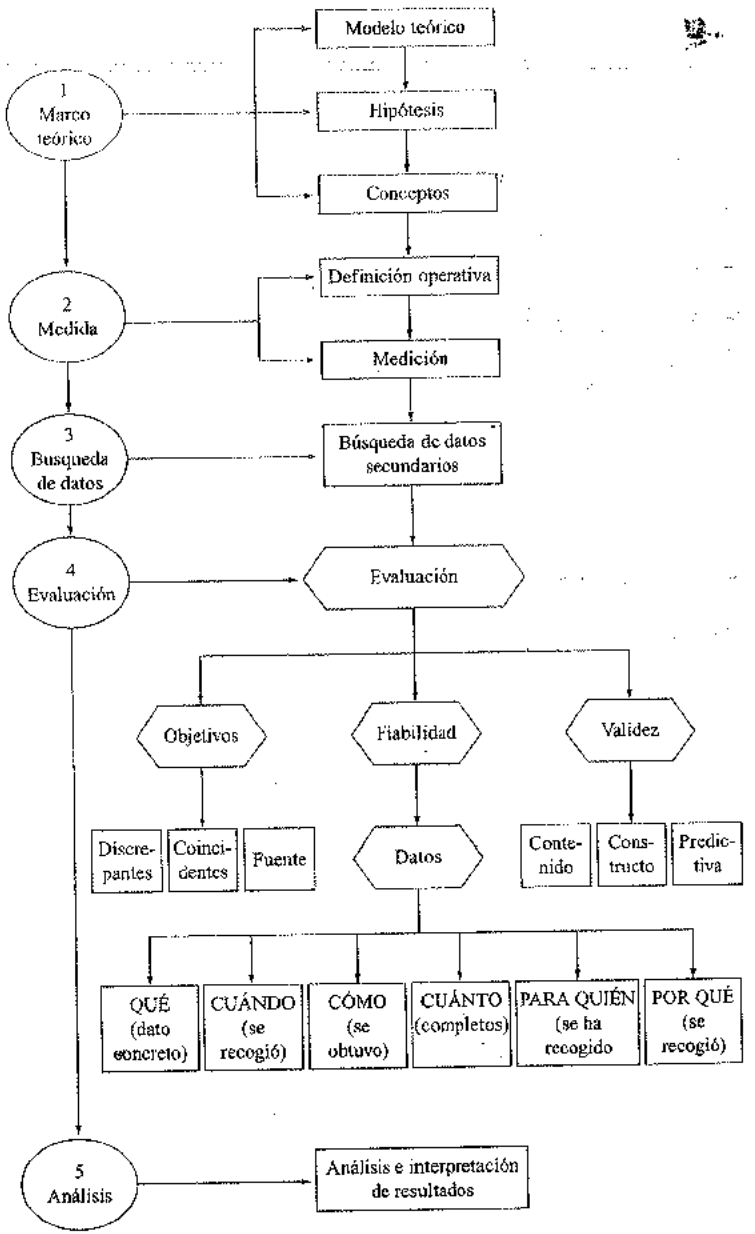
8. Guía básica para realizar una investigación basada en datos secundarios

La proliferación de información disponible no soluciona por sí misma los problemas relativos a la fiabilidad, validez y adecuación de la misma, por lo que hemos de seguir planteándonos las preguntas clásicas de la investigación. Resumimos seguidamente las preguntas que deben plantearse, formulándolas en torno a cuatro estadios o fases de una investigación basada en datos secundarios (diagrama 1). El apartado 1 se refiere al «qué» de la investigación, es decir al marco o modelo teórico que debe guiar cualquier actividad investigadora, incluyendo aquella que se basa en datos secundarios. Por su parte, el apartado 2 tiene que ver con la medición, es decir, con cómo traducimos nuestros conceptos o variables teóricas en indicadores empíricos. El apartado 3 se refiere concretamente a lo más específico de la estrategia metodológica de la investigación basada en datos secundarios: actividades de búsqueda, evaluación y selección de los datos existentes. Si los datos, finalmente, superan las pruebas de evaluación, estaremos en disposición de acometer las tareas contenidas en el apartado 4, referentes al análisis e interpretación de los resultados. Con fines expositivos, plantearemos el contenido de estos cuatro apartados en forma de interrogantes a los que, de una u otra forma, debe contestar el investigador que se adentra en una investigación basada en información/datos secundarios.

1. Marco teórico

1. ¿Está clara la idea de cuál es el problema concreto de investigación? ¿Qué es lo que se quiere saber? ¿Podría expresarse por escrito en cinco

Diagrama 1. Marco-guía para afrontar las preguntas específicas de una investigación basada en datos secundarios



líneas de forma que alguien que no conociera el tema se enterara con sólo leer este escrito?

2. ¿Se podrían aislar las hipótesis más importantes de la investigación? ¿Se pueden formular explícitamente? ¿Qué tipo de hipótesis son?
3. ¿Cuáles son los conceptos que las hipótesis tratan de relacionar? ¿Se ha preparado un listado de los conceptos que entran a formar parte de las hipótesis? ¿Se ha definido adecuadamente cada uno de dichos conceptos? Por ejemplo, ¿qué es una familia monoparental? ¿Qué significa la pregunta de un censo *cuántas habitaciones tiene su casa*? ¿Qué significa habitación? ¿Se incluye el salón-comedor o no? ¿y un cuarto de baño? ¿Qué es una familia? ¿Se incluyen los padres de los padres o no? No hay que olvidar que cuanto más abstractos sean los conceptos a medir más definiciones habrá que utilizar, para que se capte bien el carácter multidimensional de esos conceptos.

2. La medida

El investigador de datos secundarios tiene que reconstruir el proceso que generó los datos originalmente: cómo, cuándo, por quién fueron recogidos los datos. El determinar la autenticidad es un problema muy técnico. Para aumentar la fiabilidad conviene utilizar más de una fuente de información. Las preguntas siguientes pueden orientar el proceso.

1. ¿Cuáles son los indicadores empíricos que se piensa utilizar para medir las distintas dimensiones y subdimensiones de los conceptos? Antes de buscar datos secundarios debería contarse con un sistema de indicadores o variables empíricas adecuadas.
2. ¿Cómo tendrían que hacerse las medidas para que resulten fiables y válidas?

3. Búsqueda y evaluación

Esta fase es la más específica de la investigación basada en datos secundarios y debe prestársele especial atención, puesto que el investigador no controla la recogida de datos. No hay que olvidar que el éxito o fracaso de una investigación depende, en gran medida, del cuidado que se ponga en recoger datos serios, sólidos y dignos de confianza. De poco sirve hacer análisis estadísticos complejos por medio de técnicas multivariantes si el *input* es malo, si los datos no son fiables. Por estas razones, conviene que nos hagamos una serie de preguntas a la hora de buscar la información secundaria que necesitamos.

1. *Fiabilidad*

1.1. ¿Cómo y dónde se puede localizar la información necesaria? ¿Se cuenta con una buena «familiaridad» con los organismos públicos y privados que generan y almacenan datos que puedan ser de utilidad para el estudio? ¿Son accesibles estos datos? ¿Pueden obtenerse en soporte magnético? ¿En cuánto tiempo? ¿Qué información de apoyo se necesita? ¿Se puede determinar la autoridad, objetividad y precisión de la información obtenida vía Internet? Debe recordarse que si se trata de una encuesta, hará falta una copia del cuestionario, la descripción del fichero, el plan de muestreo, etc.

1.2. Ante cualquier estudio/fuente de información debemos conocer cuál fue el objetivo del mismo, porque puede suceder que su objetivo sea o coincidente o discrepante con el de nuestro estudio. Si es coincidente, podremos encontrar en él indicadores empíricos (datos) para nuestro propio estudio. Si los objetivos fueran muy discrepantes, lo más probable es que dicho estudio o fuente de información sirva de muy poco. Pero, además, el conocer el objetivo del estudio originario puede ayudarnos, en determinadas situaciones, a evaluar indirectamente la fiabilidad del mismo. Es de sobra conocido el hecho de que algunos estudios de evaluación pueden pretender acopiar datos sobre los que justificar modos de hacer o de prestar servicios de la institución.

1.3. ¿De quién fue la responsabilidad del estudio? ¿Cuál es el grado de fiabilidad de la fuente a utilizar? ¿Está claro quién es la organización responsable de la información? ¿Se mencionan explícitamente las fuentes de donde proviene la información? En la actualidad, estas dos últimas preguntas son centrales, ya que existen montones de páginas *web* cuya procedencia puede ir de desconocida a incierta. Unas fuentes son más creíbles que otras. Hay fuentes de paternidad más que dudosa. Los criterios a utilizar son variados, e incluyen, por ejemplo, la experiencia en actividades de recogida de información, la calidad de la red de campo en las encuestas, el método de recogida de datos, los mecanismos de depuración de datos, la codificación, etc.

1.4. ¿Cuál es la fiabilidad de los datos generados por la fuente en cuestión? Esta pregunta debe ser específica para cada uno de los datos que se vaya a utilizar. La evaluación directa de la fiabilidad puede resultar muy difícil en el caso de los datos secundarios, por lo que será conveniente utilizar los mecanismos de control que se tengan a mano.

1.5. ¿Qué datos concretos se recogieron? Por ejemplo, si se está trabajando sobre datos de turismo basados en el número de turistas que

visitan en un año un país, hay que saber cómo se contaron los turistas. ¿Deberíamos considerar *turista* a toda persona que haya cruzado la frontera, al margen de que lo haga por razones de trabajo o de vacaciones? ¿Entendemos por turista a personas distintas o a todos los viajeros, independientemente del número de veces que hayan entrado en dicho país?

1.6. ¿Cuándo se recogieron los datos? El factor tiempo es decisivo de cara a la persistencia del valor de los datos. Las actitudes cambian en función de circunstancias diversas, mientras que otras características son más estables.

1.7. ¿Cómo se obtuvo la información? Esta pregunta hace referencia a la metodología o procedimiento utilizado en la recogida de la información. Conviene saber, por ejemplo, el tipo de muestreo, el tamaño de la muestra, el error muestral, el método de administración del cuestionario (personal, telefónico, por correo, en grupo, individual), etc. Por otra parte, si se trata de datos obtenidos con muestras pequeñas, hay que plantearse la credibilidad de las mismas. Evidentemente, la credibilidad no es una cuestión de «todo o nada», por lo que debe valorarse la credibilidad relativa de la muestra cuyos datos pensamos utilizar[2].

1.8. ¿Cuán completa es la información recogida? ¿Cuál fue la tasa de respuesta? ¿Cuántos datos faltan? ¿Cuál fue el nivel de cobertura y características de la no respuesta? ¿Qué ítems o preguntas importantes faltan en el cuestionario original?

1.9. ¿Quién era el destinatario de la información? Ya mencionamos anteriormente el ejemplo de los datos de determinados estudios de evaluación que, al estar hechos para respaldar la política o modos de hacer de una institución, pueden haber sido manipulados convenientemente. Según Starr (1987, p. 38), *más que de falsificaciones intencionadas se trata de técnicas más corrientes que incluyen una utilización defectuosa de las clasificaciones y una tolerancia de fallos metodológicos que producen datos que pueden tener efectos políticos beneficiosos*.

1.10. ¿Por qué se recogió esa información? Los objetivos de la recogida de datos pueden ser diversos. No es lo mismo que la información se haya recogido para un fin que para otro. El por qué y el para qué son preguntas que también se hacen los entrevistados y entrevistadas ante determinadas encuestas.

2. *Validez*

Estas son algunas preguntas que hay que hacerse con respecto a los datos (indicadores o medidas) que se estén utilizando.

2.1. ¿Tienen los indicadores contenidos en los datos secundarios validez? En esencia se trata de valorar si los indicadores miden lo que se supone que miden. No es fácil valorar la validez, pero conviene intentarlo por todos los medios.

2.2. Según señala Jacob (1984), cuando no hay correspondencia entre el concepto abstracto y la medida concreta se produce el error (falta de validez de constructo). La pregunta es pues: ¿poseen los indicadores a utilizar validez de constructo? Traducida al usuario de datos secundarios esta pregunta sería: ¿Se corresponde la definición del concepto del autor de los datos secundarios con su definición de ese concepto? Expresado en términos concretos ¿coincide nuestra idea de *familia* con la contenida en el censo?

2.3. Si se están utilizando series temporales hay que preguntarse si el concepto que queremos medir sigue siendo el mismo o ha variado su significado con el paso del tiempo.

2.4. ¿Predice adecuadamente la medida que se está utilizando lo que se pretende? (validez predictiva).

2.5. Una pregunta más tendría que ver con la denominada validez de contenido desde un doble punto de vista. En primer término, ¿cubre la medida las dimensiones más relevantes del concepto? Carmines y Zeller toman el concepto de «alienación» tal como lo define Seeman (1959) en base a la sensación de impotencia, ausencia de normas, carencia de sentido, aislamiento social y autoextrañamiento. Por otra parte, el investigador de datos secundarios debe interrogarse sobre si el concepto que va a utilizar tomándolo de otras fuentes cubre o no las dimensiones relevantes que se propone manejar en su propia investigación.

4. Análisis e interpretación de resultados

Una vez que, tras la evaluación de fuentes y datos, el investigador se decide a utilizar la información recogida, comienza el análisis. La estrategia analítica depende lógicamente del diseño concreto de investigación. Planteamos seguidamente algunas preguntas-guía sobre aspectos prácticos a tener en cuenta a la hora de iniciar el análisis de los datos recogidos y depurados.

1. ¿Está claro el tipo de variables que se va a utilizar en el análisis? Un aspecto importante del análisis de datos secundarios reside en que podemos combinar la información de modo que nos permita obtener nuevas variables a partir de las existentes y representar así un nuevo concepto que no estaría en los datos originales. Es quizás en esta fase de la investigación basada en datos secundarios donde entra de lleno el ingenio y la creatividad del investigador para construir en base a los

datos que tiene (primarios) nuevos conceptos y operacionalizaciones de los mismos.

2. ¿Está claro y se maneja adecuadamente el programa estadístico oportuno para el análisis de las variables? Existen en la actualidad muchas posibilidades al respecto: SPSS, BMDP, SAS, BARBRO, MINITAB, etc. Por otra parte, ¿es necesario combinar distintos tipos de programas? ¿Existen procedimientos analíticos (programas) que permitan construir modelos de un modo eficiente y prácticos?
3. Si ya está preparado el fichero de datos, no olvidar, antes de empezar el análisis de datos como tal, verificar la información, recurriendo a distintos mecanismos como, por ejemplo, (1) obtener una distribución de frecuencias de aquellas variables para las que se disponga de información por otras vías para ver si ambas distribuciones concuerdan. Si se está trabajando con datos censales, puede tomarse la distribución de edades de cualquier publicación del INE y compararla con la misma distribución de los datos que has obtenido (Dale *et al.* 1988, p. 135); (2) hacer un diagrama de dispersión para ver si existen casos que se desvían excesivamente de la pauta general y que resulten inexplicables a partir de la teoría que se está utilizando.
4. Si los datos provienen de una muestra, ¿se ha calculado o comprobado el error de muestreo existente? Conviene no olvidar que en España los informes —publicados o no— no suelen incluir el cálculo de los errores de muestreo, para las principales variables al menos.
5. ¿Se ha comprobado la unidad de análisis de los datos? No es lo mismo una encuesta a un fumador individual que una encuesta a ese fumador como informante de su hábito en sí mismo y en todos los miembros del hogar. No siempre es cierto que unidad de análisis y unidad de observación coincidan. En este sentido, el nivel de análisis puede plantearse a distintos niveles: hogar, unidad familiar, individuo. Téngase en cuenta que es aquí donde suele presentarse el riesgo de cometer la denominada falacia ecológica.
6. Finalmente, el cuadro 1 puede utilizarse como un prontuario para detectar posibles problemas y soluciones en los datos que estemos manejando desde el punto de vista de su fiabilidad y validez.

2. Fuentes de datos disponibles

Cada vez van siendo más numerosas y extensas las fuentes de datos utilizables por los científicos sociales. Debemos comenzar señalando que la simple enumeración de estas bases de datos excedería los límites de este trabajo. Lo único que pretendemos es reseñar algunos ejemplos de fuentes de datos estadísticos secundarios en sociología, sin ningún ánimo de exhaustividad. Las fuentes que se citan tratan sólo de ejemplificar. No resulta fácil su clasifica-

Cuadro 1. Resumen de los problemas de los datos y posibles soluciones

Problema	Solución
ERRORES DE SELECCION	• En el caso de muestreo aleatorio, determinar el error de muestreo. • Para los recuentos y muestras no aleatorias, redondear para advertir el error y evitar dar la impresión de precisión exagerada.
NO-VALIDEZ: Validez de constructo: Falta de ajuste entre la conceptualización del recolector de los datos y del usuario.	• Para diagnosticarla, buscar la convergencia con otras medidas y/o la capacidad de discriminar con respecto a otros conceptos. • Seguidamente, utilizar conceptos múltiples, elegir la medida alternativa más adecuada, o abandonar el estudio si no existe una medida válida.
Los cambios de situación producen invalidez.	• Buscar y utilizar la variable / indicador común subyacente tanto al fenómeno pasado como al actual. • Calcular el factor de «conversión» para pasar del antiguo al nuevo. • Realizar análisis separados para los períodos de tiempo en los que la definición permanecía constante.
Transformaciones inadecuadas.	• Utilizar la variable adecuada para transformar los datos. • Vigilar los errores que pueda contener la variable utilizada para hacer la transformación. • Tener en cuenta la invalidez potencial de la variable transformadora. • Evitar utilizar la misma transformación tanto para la variable dependiente como la independiente.
No disponibilidad de los datos necesarios para los momentos necesarios.	• Interpolar utilizando métodos log-lineales, complementados con datos adicionales.
FIABILIDAD: Errores manuales.	• Buscar los casos «desviados» por medio de un diagrama de dispersión. • Verificar el grado de formación que dan los organismos de recogida de datos. • Redondear.
Cambios en los procedimientos de recogida de datos.	• Hacer evaluaciones distintas de la fiabilidad para cada uno de los segmentos de datos.
Correcciones de los datos hechas por los organismos.	• Buscar información sobre los procedimientos de corrección utilizados por el organismo correspondiente y en qué momentos se aplicaron esas correcciones.

Cuadro 1. Resumen de los problemas de los datos y posibles soluciones (continuación)

Problema	Solución
Manipulación de los datos.	• Buscar información sobre dichas manipulaciones. • Hablar con personas de dentro de la organización.
Instrumentación.	• Hacerse con una copia del instrumento de recogida de datos y evaluarlo en función de los errores de instrumentación.
Categorización.	• Localizar las inconsistencias temporales y espaciales. • Intentar categorizar los datos en categorías más coherentes.

FUENTE: Esta tabla esta tomada de Jacob (1984), *Using published data: Errors and remedies* (Londres: Sage 1984), pp. 47-48.

ción sistemática, sobre todo porque carecemos de guías adecuadas de dichas fuentes y las que existen, en forma de catálogos de publicaciones, son poco comprehensivas y están más bien dispersas. Sería esta tarea de recopilación enormemente fructífera, necesaria y útil para los investigadores sociales. En este sentido, habría que escribir de nuevo un libro como el publicado en 1969 por Amparo Almarcha *et. al., La documentación y organización de los datos en la investigación sociológica,* con los mismos objetivos pero adaptándolos a la nueva realidad de los ordenadores y las bases de datos informatizadas que hoy existen y entonces no. La importancia de un esfuerzo de este tipo queda bien reflejada en la «Presentación» del citado libro:

> No es nuestra intención en este caso el construir aquí un nuevo manual de métodos y técnicas sociológicas, sino de proporcionar una herramienta previa y no menos útil, pensando sobre todo en la situación actual de la investigación social en nuestro país. En efecto, antes de poder utilizar la metodología de la investigación sociológica propiamente dicha, los sociólogos que desean trabajar en España necesitan orientarse sobre la problemática general de cómo organizar el material o los datos, cómo recoger uno u otro tipo de información y qué hacer con ella. Las páginas que siguen tratan de limpiar el camino que conduce a la resolución de esa problemática, según nuestra experiencia mucho más compleja de lo que a veces suponen los investigadores bisoños (p. 7).

1. Documentos y estadísticas oficiales españolas

La mayoría de los documentos escritos son públicos. Al hablar aquí de *documentos* nos referimos a dictámenes, sentencias judiciales, legislación, ordenanzas municipales, certificados de nacimiento, matrimonio y defunción, censos, directorios, almanaques, anuarios, memorias anuales, boletines esta-