SECRETARIADO TECNICO DE LA PRESIDENCIA
OFICINA NACIONAL DE ESTADISTICA

VIII CENSO NACIONAL DE POBLACION Y VIVIENDA

Evaluación del proceso de entrada de datos
Informe final
(Borrador)

Elaborado por:
Julio César Mejía
Demógrafo, PhD©
Asesor

Febrero de 2004

PERSONAL QUE PARTICIPÓ EN LA EVALUACIÓN

Coordinación
Julio César Mejía

Diseño de la muetra:
José Achécar Chupani,

Apoyo informático
Bolívar Gil
Rahnsés Marquez
Guillermo Molineaux

Personal verificador
Campos de marcas y número:
Golsky M. Reynoso
Neira E. Pérez
José Miguel Fernández
Antonia Castillo
Jesús Díaz Gell
Ramona Morel

Campos de codificación automática
Rafaela Rocha
Urba Díaz
Fulvio Batista
Carmen Julia Mejía
Australia Cuevas

1. Objetivos de la evaluación

Para el proceso de captura y almacenamiento de los datos censales la ONE contrató a la firma INVERSIONES MACRO S.A. (DATOCENTRO), que a su vez contrató a la firma Ingenieros Consultores y Asociados (ICA), con asiento en Uruguay, responsable de la captura de los datos en los censos de Uruguay de 1998 y de Chile en el 2001.

En la captura de los datos censales se utilizó la tecnología de imágenes electrónicas, que son tomadas a las boletas censales y convertidas a su vez en datos en formato ASCII (*American Standard Code for Information Interchange*). Un equipo conjunto de ONE y la firma contratada controló la calidad de la lectura durante todo el proceso mediante el examen de muestras sistemáticas de boletas de cada uno de los lotes de trabajo[1] entregados por la ONE a la empresa contratada

No obstante, la ONE ha considerado pertinente una evaluación final de todo el proceso de reconocimiento óptico, incluido una verificación a una muestra de boletas con independencia de la firma contratada que capturó los datos censales que permitiese verificar que las tasas de reconocimiento y de error en los datos entregados estuviesen acorde con los límites de error aceptados en los términos contratados.

2. Características de la tecnología de captura de datos utilizada

Aunque aplicada por primera vez en la República Dominicana, la metodología de procesamiento óptico utilizada ha sido construida basándose en la experiencia adquirida en procesos reales, evolucionando progresivamente. Esta novedosa modalidad de captura utiliza escáner de reconocimiento óptico de marcas y textos alfanuméricos escritos manualmente mediante el software Sistema Integrado de Procesamiento Óptico (SIPO), adaptado a las características propias del país.

2.1 Especificaciones técnicas de los escáneres utilizados

Los escáneres usados son de marca Fujitsu, modelo 4099D, el cual posee una velocidad de 180 páginas por minuto en modo duplex (anverso y reverso e cada hoja) en forma simultánea a 200 dpi, escaneando documentos formato A4 o carta (8.5 x 11 pulgadas) en modo horizontal.

Con respecto a su mantenimiento, esos escáneres están diseñado para funcionar durante 8 horas o más sin mantenimiento, funcionando en la práctica hasta 24 horas ininterrumpidamente, con paradas de mantenimiento diario de menos de una hora, y posee mecanismos de acceso muy simples de acceso a todos los componentes internos que requieren ser limpiados diariamente para eliminar los residuos de papel, goma de borrar u otras pequeños residuos.

2.2 Características del software utilizado en el reconocimiento óptico

El Sistema Integrado de Procesamiento Óptico (SIPO) utiliza simultáneamente tres motores de reconocimiento de los más potentes del mercado: Nestor Reader (NCS-

[1] Unidas operativas de agrupamiento de 700 a 800 boletas para su procesamiento óptico.

EEUU), Kadmos (Reconognition-Alemania) y Fineeader (Abbys-Rusia). Estos motores poseen la capacidad de reconocer campos de marcas (OMR), campos de texto preimpreso (OCR), campos de texto manuscrito (ICR) y campos impresos de códigos de barras.

El SIPO también posee un módulo que tiene la función de transformar los textos obtenidos del procesamiento óptico de los campos de ocupación, rama de actividad, y otras respuestas alfabéticas en valores numéricos correspondientes a códigos. Para alcanzar este resultado, el módulo trabaja en dos fases: fase automática y fase de codificación asistida, en forma interactiva.

La fase automática consiste en utilizar diccionarios de textos, buscando en ellos cada texto reconocido. Cuando el texto existe en el diccionario, se asocia el código correspondiente. En la fase asistida, el software presenta en pantalla los textos que no están en los diccionarios de códigos, permitiendo que un operador especializado ingrese el valor del código, asistiéndolo con información descriptiva de los códigos.

Una vez realizado el reconocimiento óptico de los campos alfabéticos, éstos son analizados por un módulo de Reglas de Integridad Semánticas, donde se busca reconstruir los textos incompletos. Para ello se utiliza información proveniente de diccionarios específicos, asociados a la semántica de cada campo. Estos diccionarios permiten el uso de algoritmos de análisis de texto, donde se busca la similitud entre el texto reconocido y los diccionarios. Estos algoritmos permiten reconstruir palabras o campos a partir de sólo una parte del texto bien reconocido.

Finalmente, el SIPO posee mecanismos avanzados para la corrección automática de caracteres, tanto para los campos alfabéticos como numéricos, mediante la aplicación de Reglas de Integridad cuando la inconsistencia es producto de un error en el reconocimiento óptico.

3. Aspectos metodológicos sobre la evaluación

La evaluación se realizó en dos partes: la primera se basó en los reportes finales de control de calidad entregados por la firma contratada, y la segunda, realizada después de haber concluido el proceso de captura (a posteriori), consistió en un cotejo o verificación de las informaciones contenidas en una muestra aleatoria de boletas, comparando las informaciones en las imágenes digitalizadas de las boletas con los datos almacenados en formato accesible mediante el SIPO.

3.1 Sobre el proceso de captura de los datos

Previo al escaneo de las boletas de cada lote, equipos de trabajo de la ONE conformados por 5 revisores y un supervisor revisaban y/o corregían las informaciones contenidas en la sección de identificación geográfica, o sea, el número de la carpeta, número de la vivienda, número del hogar, así como el Resumen de Población, el número de la persona, la relación o parentesco con el jefe del hogar (p27), el sexo (p28), fecha de nacimiento (p29) y edad (P30).

La codificación de las preguntas abiertas o variables alfanuméricas (ocupación, rama de actividad, lugar de nacimiento, residencia 5 años antes, país de nacimiento del

padre y país de nacimiento de la madre) se realizó de manera automática en alrededor del 80% de los casos mediante el SIPO, utilizando las tablas de codificación elaboradas por la ONE, incluido el sistema de codificación geográfica, adaptando para ello diccionarios y códigos de ocupaciones, ramas de actividades, carrera y postgrado utilizados en censos de otros países, basados en sistemas clasificatorios internacionales. Esta codificación automática fue complementada con la codificación asistida por un operador de verificación manual de la ONE.

3.2 Acerca de los procesos de control de calidad e integridad

El SIPO posee mecanismos de verificación y acciones asociadas que garanticen la calidad de los datos procesados, en términos e su integridad semántica. Esta consistencia se refiere a las respuestas registradas en las boletas censales. Estos mecanismos consisten en la definición de Reglas de Integridad Semántica, que son definidas en SIPO y evaluadas en tiempo real durante el proceso de reconocimiento, reglas que permiten obtener un conjunto de boletas más coherentes, y de mejorar la calidad del reconocimiento óptico mediante la validación de propiedades lógicas.

También dispone el SIPO de dos módulos complementarios para la evaluación de la integridad de los datos: el módulo Reglas de Verificación (quién actúa en forma temprana, detectando eventuales errores de reconocimiento que generan violación de Reglas de Integridad de los datos) y el módulo de Control de Integridad de Lotes. En ambos casos se trata de módulos interactivos, donde un operador especializado puede verificar o corregir datos de reconocimiento.

3.3 Sobre la evaluación a posteriori

3.3.1 La verificación

La verificación se realizó comparando la información almacenada en la imagen óptica de cada boleta para cada uno de los campos seleccionados con el dato almacenado en formato ASCII. Para ello se seleccionó una muestra de 146 carpetas y 1,500 boletas en 20 provincias, de un total de 32 provincias y un universo de cerca de 2.3 millones boletas. Los registros incluyen además de los datos capturados automáticamente (sin intervención humana), las informaciones que fueron revisadas manualmente, hayan sido o no modificadas por el operador o revisor.

A la muestra de boletas seleccionadas se le verificó la sección de identificación (sección I) completa (número de la carpeta, número de la vivienda y número del hogar dentro de la vivienda), así como una muestra de 6 preguntas de marcas, 9 de caracteres numéricos, y 7 alfabéticas correspondientes a las secciones II (características de la vivienda), III (Identificación de los hogares en la vivienda), IV (características del hogar) y VII (características personales).

Para el registro de los resultados se diseñó un formulario (anexo) para cada boleta examinada. El resultado del cotejo en cada campo con información contenida en la boleta se clasificó como sigue: reconocimiento correcto (C), reconocimiento incorrecto (I), no reconocimiento (E).

3.3.2 Descripción metodológica de la muestra

El principal criterio para determinar el tamaño y la selección de la muestra se fundamentó en el tiempo disponible para la revisión de las mismas y el número de personas que realizaría dicha labor, teniendo en cuenta el objetivo de la verificación del proceso de captura de imágenes y datos. En consecuencia, el equipo técnico del VIII Censo Nacional de Población y Vivienda se permitió tomar decisiones sobre el número de carpetas, boletas y personas a ser seleccionadas, que no se corresponden estrictamente con la formalidad que conlleva un diseño muestral propiamente dicho. En este sentido, se acordó la selección de un total de 1,500 boletas. El número total de boletas revisadas, o sea, con información en los campos seleccionados fue de 1,583.

El total de provincias se dividió en dos grandes estratos. El primero, conformado por las provincias de mayor tamaño, que concentran el 80% del total de viviendas del país, de inclusión obligatoria (IO), o sea, con probabilidad de selección igual a la unidad, mientras que el segundo corresponde al resto de las provincias (OP). Las carpetas de las provincias fueron seleccionadas al azar y de estas se escogieron al azar la mitad de ellas. El total de hogares seleccionados en este grupo fue de 1,401.

Respecto al estrato OP, el equipo técnico de la ONE, determinó escoger cinco (5) provincias y de estas se procedió, al igual que en el grupo anterior, a la selección aleatoria de las carpetas, y de éstas la mitad de las boletas, también de manera aleatoria. El total de boletas seleccionadas en este estrato fue de 182.

Para la verificación de las informaciones correspondientes a las personas, se decidió realizar una selección en otra etapa de un máximo de cuatro personas, incluido el jefe del hogar, si el total de personas residentes en el hogar superase dicho número.

4. Resultados de la evaluación

4.1 Los reportes de controles de calidad

Al final del proceso la firma contratada entregó a la ONE reportes de la evaluación del procesamiento óptico de las boletas censales, resultados de los controles de calidad implementados, basados en tomas sucesivas de muestras. Los tamaños de muestra para cada una de los campos varían significativamente. Al final del proceso, los tamaños de muestra van desde 11,5% para los años de estudios de la primaria hasta 99.4% en el caso del año de nacimiento. De los 37 campos de caracteres numéricos evaluados, en 26 de ellos (70%), los tamaños de muestras superan el 30% del total de casos leídos , y en 15, o sea, el 40.5%, la muestra es superior al 80% de los casos leídos.

En promedio los campos de marcas fueron reconocidos en un 100%, los numéricos en un 95% y los campos de caracteres alfabéticos en un 75.5%. Este relativamente baja tasa de reconocimiento se debe sobre todo a errores en la escritura de los literales o textos (de ortografía o sintaxis), lo que obligó a enviar a Verificación Manual para su corrección el 40.4% de los campos con información alfabéticas. Los porcentajes de errores de reconocimiento de marcas, números y caracteres alfabéticos fueron muy bajos: 0.47, 0.25 y 0.71 por ciento respectivamente.

Al finalizar del proceso, el porcentaje de campos numéricos correctamente reconocidos varía desde 77.5% (hermanas embarazadas fallecidas por embarazo, parto o puerperio) hasta 99.2% (persona no.). En 20 de los campos (54.1%) los casos correctamente reconocidos superan el 95%, y en 11 (29.7%) se reconoció correctamente entre 90% y menos de 95%, y en el 83.8% de los campos se reconoció más del 90%. Los demás campos con los más bajos porcentajes de reconocimiento fueron ingreso mensual (86.1%), año de llegada al país (86.3%), cantidad de dinero recibido del exterior (88.4%), año de nacimiento del último hijo o hija nacido vivo (89.4%). Los campos con mayores porcentajes de reconocimiento automático fueron el número del hogar (98.5%), No. de la carpeta (98.0%).

Por consiguiente, los campos de números presentan muy bajas tasas de error: el 56.3% tuvieron menos de 1% de casos con error, 34.3% de los campos leídos con información de 1% a menos de 3%, y sólo el 10% restante tuvo márgenes de error 4% a 4.8%. Los campos numéricos mayores porcentajes de errores de reconocimiento fueron años de estudio de doctorado (5.6%), años de secundario (4.8%), años de primaria (4.47%), los años universitario (4.22%), años de especialidad (4%), años de maestría (2.54%), años de preescolar (2.54%). Los campos con menores errores de reconocimiento fueron, además del número del hogar, el número de la carpeta (0.3%), el año de nacimiento (0.3%), número de varones fuera del país (0.4%), hijos actualmente vivos (0.4%), hijas actualmente vivas (0.5%), y edad (0.6%).

En general las tasas de errores encontradas son bastante parecidas a las tasas de error estimadas (error residual).

Cuadro 1
Indicadores de reconocimiento y error en la lectura según tipo de campo
(Porcentajes)

Nombre del campo o variable	Reconoc. Auto [2]	Verificación Manual [3]	Error residual[4]
Marcas	100.0	0.9	0.2
Números	95.2	17.9	0.3
Alfabéticos	75.5	40.4	0.7

Respecto de los campos de respuestas abiertas codificadas los porcentajes de reconocimiento automático fluctúan entre 48% (postgrado) y 68% (País de nacimiento de la madre), y los porcentajes de codificación automática van de 75.6% en la rama de actividad a 95.8% en la carrera básica. Los altos porcentajes de campos enviados a verificación manual se debe sobre todo a errores en la escritura de los literales o textos. De todas maneras, las tasas de error de este tipo de campo al final del proceso automático y manual son sólo un poco más elevadas que no los campos de números, pero inferiores al 4% en todos los casos. Las preguntas sobre el lugar de nacimiento y el lugar de residencia 5 años antes presentaron las mayores tasas de errores (3.4% y 3.6% respectivamente).

[2] **Reconocimiento automático**: Campos reconocidos y procesados automáticamente (sin intervención humana).
[3] **Verificación Manual**: Porcentaje de campos (con relación al total de campos de ese tipo que contiene información en las boletas) que han sido enviados al proceso de Verificación Manual.
[4] **Error residual**: Porcentaje estimado de campos con error al momento de finalizar el procesamiento óptico

Cuadro 2
Indicadores de reconocimiento y error en la lectura de campos numéricos
(Porcentajes)

Nombre del campo o variable muestra[9]	Reconoc. Auto[5]	Reconoc. real[6]	Error	Error [7]	Tamaño residual[8]
Número de carpeta	97.9	98.0	0.3	0.2	38.3
No. de la vivienda	85.6	96.6	0.7	0.0	99.4
No. del hogar	93.7	98.5	0.2	0.0	99.4
Persona No.	89.1	99.2	0.7	0.6	16.0
Persona No.	79.2	98.5	1.1	0.9	15.6
No. cuartos	82.3	96.9	0.7	0.1	80.6
Varones fuera del país	89.1	95.0	0.4	0.1	72.9
Hembras fuera del país	88.6	95.0	0.4	0.1	73.4
Cantidad de dinero del exterior	64.8	88.4	1.4	1.2	16.5
Edad	83.5	94.3	0.6	0.0	99.4
Fecha de nacimiento					
Día	85.3	94.1	.7	1.2	8.1
Mes	79.5	95.0	1.2	0.9	29.0
Año	74.7	92.3	0.3	0.0	99.3
Año de llegada al país	72.5	86.3	1.4	1.0	28.6
Años de estudios					
Preescolar	89.0	95.7	2.5	2.2	14.2
Primaria	85.6	93.7	4.5	4.0	11.5
Secundaria	86.6	93.2	4.8	4.0	17.6
Universitaria	86.7	93.4	4.2	2.9	30.8
No. de empleados	79.4	92.4	1.6	1.1	31.3
Ingreso mensual	60.0	86.1	1.4	1.2	16.2
No. de hijos	77.6	97.2	0.7	0.1	80.3
No. de hijas	79.7	97.2	0.8	0.2	80.6
Hijas actualmente vivas	79.1	96.9	0.5	0.1	80.4
Hijos actualmente vivos	76.3	96.8	0.4	0.1	28.0
Año de nacimiento ultimo hijo	64.6	89.5	1.1	0.8	29.3
Mes de nacimiento último hijo	85.4	94.6	1.2	0.9	28.0
No. de hermanas	79.7	97.5	0.7	0.1	80.4
No. de hermanas fallecidas	81.4	97.2	0.5	0.1	78.6
No. de fallecidas por embarazo	63.8	77.5	2.5	0.9	63.4
Total de personas	83.4	95.9	1.1	0.0	99.4
Total de varones	83.7	96.5	0.8	0.0	99.3
Total de hembras	83.7	96.3	0.7	0.1	99.4
Total de personas de 18 años y más	76.7	92.0	0.3	0.0	99.4

[5] Porcentaje de campos reconocidos y procesados de manera correcta automáticamente (sin intervención humana) + campos enviados a verificación manual y que estaban correctamente reconocidos.
[6] Porcentaje de campos reconocidos al final del proceso. Incluye campos reconocidos automáticamente + campos corregidos por el operador de verificación
[7] Porcentaje de reconocimiento incorrecto
[8] Porcentaje de campos estimados con error al momento de finalizar el procesamiento óptico
[9] Porcentaje que representa la muestra con respecto al total de campos leídos

Fuente: Informe de calidad y reconocimiento de ICA

Cuadro 3
Indicadores de reconocimiento y error en la lectura de campos de respuestas abiertas
(Porcentajes)

Nombre del campo residual [14]	Reconoc. Auto[10]	Verificación. [11]	con error [12]	Error	Error [13]
Ocupación	51	19	23	0.2	1.2
Rama de actividad	53	17	30	1.6	1.6
Carrera básica	51	20	29	1.7	1.7
Postgrado	48	20	32	2.5	2.5
Lugar de nacimiento	65	15	20	3.4	3.4
Residencia 5 años antes	65	13	22	3.6	3.6
País nacimiento de la madre	68	9	22	2.8	2.8
País nacimiento del padre	67	11	22	2.9	2.9

Fuente: Reporte gráficos de ICA

Cuadro 4

% de codificación automática de las preguntas abiertas

Preguntas	%
Lugar de nacimiento	88.9
Ocupación	78.5
Rama de actividad	75.8
Carrera	95.8

Fuente: Reporte de codificación automática de ICA

4.2 La evaluación a posteriori

[10] Campos reconocidos y procesados automáticamente (sin intervención humana).

[11] Campos enviados a Verificación manual, pero no modificados por el operador, ya que habían sido in correctamente reconocidos por el software

[12] Campos enviados a Verificación manual y modificados por el operador, ya que estaban correctamente reconocidos por el software de manera automática

Reconocimiento incorrecto

[13] Campos reconocidos incorrectamente o no reconocidos al momento de finalizar el procesamiento óptico

En general, al igual que lo reportado por la firma contratada, se encontraron altos porcentajes de reconocimiento correcto, y por consiguiente, bajas tasas de error. Los campos con respuestas de marcas presentan los mayores porcentajes de reconocimiento. Los porcentajes de reconocimiento correcto oscilan entre 95.9% y 100%. Los mayores porcentajes se encontraron en la relación o parentesco (100%), estado conyugal (100%), condición de alfabetismo (99.9%), número del hogar (99.8%), y sexo (99.7%). Los menores porcentajes de reconocimiento fueron para rama de actividad (95.9%), y ocupación (96.6%).

Las mayores tasas de error se encontraron en los campos de respuestas alfabéticas (codificadas), sobre todo en la rama de actividad, ocupación, y el lugar de residencia 5 años antes. También se encontró campos de respuestas numéricas entre aquellos con mayores tasas, sobre todo los años de estudio, el día y el mes de nacimiento.

Las tasas de error total van desde 0% (parentesco, estado conyugal, país de nacimiento de la madre) a 4.1% (Rama de actividad). En 10 de los campos examinados (40%), la tasa de error es inferior a 1%. Los campos con tasas de error total más bajas son el alfabetismo (0.1%); número del hogar (0.2%); sexo (0.3%), número de hogares (0.4%); y el ingreso (0.4%).

Cuadro 5
Reconocimiento en la lectura por tipo de campo (%)

Nombre del campo	Rec.[15] total	Rec. [16] correcto	Tasa [17] error	No.[18] casos
De caracteres numéricos				
No. de la carpeta	99.4	99.4	0.6	1,536
No. de la vivienda	99.4	99.4	0.6	1.536
No. del hogar	99.9	99.8	0.2	1,536
No. de hogares	9.6	99.6	0.4	1,536
No. de dormitorios	99.1	98.5	1.5	1,536
Edad	98.9	98.5	1.5	4,719
Fecha de nacimiento				
Día	98.8	97.2	2.8	4,719
Mes	98.5	97.5	2.5	4,719
Año	99.7	99.1	0.9	4,719
Años de estudios	98.8	97.2	2.8	4,719
Ingreso mensual	100.0	99.4	0.6	4,719
No. de hijos	99.9	97.9	3.1	4,719
De marcas				
Bienes y servicios del hogar	99.2	99.1	0.9	1,536
Relación o parentesco	100.0	100.0	0.0	4,719
Sexo	99.7	99.7	0.0	4,719

[15] Reconocimiento total: Total de campos reconocidos, ya sea correcta o incorrectamente
[16] Reconocimiento correcto: Campos reconocidos correctamente.
[17] Error: Campos no reconocidos, más campos reconocidos incorrectamente
[18] Total de campos con información que fueron verificados. Son los denominadores utilizados para el cálculo de los porcentajes de reconocimiento y la tasa de error.

Condición de alfabetismo	99.9	99.9	0.1	4,719
Nivel de estudios	99.5	99.5	0.5	4,719
Estado conyugal	100.0	100.0	0.0	4,719

Cuadro 6

Codificación de las preguntas abiertas

Preguntas	Codificación		No reconoce	No. casos
	Correcta	incorrecta		
Lugar de nacimiento	98.9	0.9	0.2	1,286
Lugar de residencia 5 años antes	98.0	1.0	1.0	295
País de nacimiento de la madre	00.0	0.0	0.0	76
Ocupación	96.6	3.4	0.0	2,315
Rama de actividad	95.9	4.1	0.0	2,133
Carrera	8.4	1.5	0.1	719

5. Conclusiones

Las informaciones contenidas en los reportes de control de calidad y de integridad del procesamiento óptico de la boletas censales, así como los resultados de la verificación a posteriori de los datos capturados basada en una muestra de boletas realizada por un equipo de la ONE indican altas tasas de reconocimiento óptico y de la codificación automática correctos, con bajos porcentajes de errores de reconocimiento (no reconocimiento o reconocimiento incorrecto de los caracteres que en ningún caso superan el 5%.

De acuerdo con los reportes de calidad e integridad de la firma contratada, el 56.3% de los campos (variables) tuvieron menos de 1% de casos con error, 34.3% de los campos leídos con información de 1% a menos de 3% , y sólo el 10% restante tuvo márgenes de error 4% a 4.8%.

Como era de esperarse, los registro de las respuestas de tipo marca muestran los mayores porcentajes de reconocimiento total y de reconocimiento correcto, seguidos por los registros de tipo numérico. Las menores tasas de error se presentan en los campos alfabéticos o respuestas abiertas de texto manuscrito.

Por otro lado, la reducción del tiempo en todo el proceso de captura ha sido significativa. Aún cuando se hace imposible establecer con precisión el tiempo en la captura de los datos de los censos anteriores al de 1981 y 1993, dada la precaria memoria censal documentada que posee la ONE, probablemente sea la primera vez que en apenas unos 8 meses se finaliza el proceso de captura, y en unos otros 10 meses el proceso de depuración de los datos y tabulación de los resultados definitivos.

La codificación automática de las preguntas abiertas, especialmente la ocupación, la rama de actividad, la carrera y postgrado evitaron las arduas tareas con numeroso personal de codificación y digitación manual durante el proceso de depuración o limpieza de los datos, en la etapa de control de estructura de los datos, así como en el análisis de las inconsistencias. Problemas que implicaban consumo de tiempo y personal en la búsqueda de boletas físicas, especialmente en la identificación de las boletas de continuación, en la identificación y corrección de las doble marcas – especialmente en el tipo de vivienda.

Otra ventaja adicional de esta tecnología es que permite a la ONE disponer por primera vez un archivo o base de imágenes de boletas fotografiadas y almacenadas en dispositivos (CD) y que pueden ser manipuladas y traídas a la pantalla de un computador personal con ayuda del SIPO, facilitando notablemente el proceso de control de calidad durante el proceso de captura, y de su depuración posterior, reduciendo además significativamente los tiempos de verificación, de detección y corrección de errores de consistencia.

No obstante el alto nivel de reconocimiento correcto de todo el proceso de captura de los datos censales en conjunto, quedan abiertas interrogantes respecto de los porcentajes de no reconocimiento automático y de reconocimiento incorrecto nada despreciables. El efecto que pudiese tener en estos márgenes de errores las exigencias o condicionantes de la tecnología en términos de la calidad de la escritura y de las marcas y de características físicas de la boleta (marcas de corte, bordes, dimensiones de los rectángulos o cuadrículas y óvalos), la calidad de los escáneres (programa de escaneo o digitalización ajustes de lectura, etc) y del trabajo de los responsables de los ajustes del mantenimiento y ajuste del sistema (administrador del SIPO, operadores de controles de calidad, operarios de los escáneres, los encargados de digitalización encargados de mantenimiento, entre otros) son una especie de caja negra. En qué medida los errores de reconocimiento, sobre todo en las respuestas textuales, se deben a problemas de ajuste del sistema de lectura de los escáneres o a deficiencias en la escritura de textos y números y/o llenados de óvalos es difícil de precisar con la información disponible.

No obstante los inconvenientes y problemas de calidad en la escritura de textos y en menor medida en las marcas y caracteres y en la escritura de números como resultado de las serias deficiencias ortográficas y caligráficas de un alto porcentaje de los estudiantes de los niveles medios y universitarios, y en menor medida de profesionales, pero en porcentaje significativo (y muy especialmente entre los educadores de enseñanza primaria y media), puede afirmarse que en términos de la correspondencia de los registros de datos almacenados y lo registrado en las boletas por los empadronadores, la tecnología de reconocimiento óptico ha resultado ser bastante eficiente.

ANEXO I

Cuadro No.1

Distribución de las carpetas y boletas seleccionadas según las provincias del estrato IO

Provincia	No. de carpetas seleccionadas	No. de boletas seleccionadas
Estrato IO		
Santo Domingo	26	272
Distrito Nacional	20	245
Santiago de los Caballeros	19	129
San Cristóbal	10	129
La Vega	8	95
Puerto Plata	7	78
San Pedro de Macorís	6	48
Duarte	6	68
La Romana	4	40
San Juan de la Maguana	4	60
Espaillat	4	42
La Altagracia	4	41
Monte Plata	4	46
Azua	4	63
Valverde	4	45
SUBTOTAL	130	1,401
Categorías OP		
Dajabón	2	23
Monseñor Nouel	5	55
Samaná	3	25
Peravia	4	55
Bahoruco	2	24
SUBTOTAL	16	182
TOTAL	146	1, 583

ANEXO II
FORMULARIOS UTILIZADOS