UNIVERSIDAD NACIONAL AUTÓNOMA DE MÉXICO

FACULTAD DE CIENCIAS

IMPLEMENTACIÓN DE LA EVALUACIÓN DE CRITERIOS DE VALIDACIÓN EN EL SISTEMA DE VALIDACIÓN DE LA ENCUESTA DE POBLACIÓN Y VIVIENDA 2015

REPORTE DE TRABAJO PROFESIONAL

QUE PARA OBTENER EL TÍTULO DE:

ACTUARIO

P R E S E N T A :

FERNANDO ESTAÑOL SALINAS

TUTORA:

DRA. ELISA VISO GUROVICH

2016

1. Datos del alumno
Estañol
Salinas
Fernando
55 32 71 10 33
Universidad Nacional Autónoma de México
Facultad de Ciencias
Actuaría
08032980-0

2. Datos del tutor
Dra.
Elisa
Viso
Gurovich

3. Datos del sinodal 1
Dra.
Amparo
López
Gaona

4. Datos del sinodal 2
M. en C.
Virginia
Abrín
Batule

5. Datos del sinodal 3
Dra.
María de la Luz
Gasca
Soto

6. Datos del sinodal 4
Dr.
Canek
Peláez
Valdés

7. Datos del trabajo escrito.
Implementación de la evaluación de Criterios de Validación en el Sistema de Validación de la Encuesta de Población y Vivienda 2015
43 p
2016

Dedicatorias

Dedicado con profundo amor, agradecimiento y admiración a mis padres, por la confianza, libertad y apoyo que en todo momento me brindaron para decidir mi camino. Por sus elocuentes enseñanzas a través del propio ejemplo, de su forma de encarar las adversidades y fortunas de la vida.

A mi amado hijo que me hace sentir muy orgulloso como padre, que ya no recibe mis consejos sino mis opiniones, pues es capaz de tomar sus propias decisiones asumiendo sus consecuencias.

A Liz con inmenso cariño, además, por su invaluable apoyo en tiempos buenos y malos.

A Natalia Volkow y Virginia Abrín, con todo mi afecto, respeto y admiración. Por su cordial insistencia que mucho influyó en mi decisión de cerrar este ciclo.

Agradecimientos

Agradezco a mi tutora, Dra. Elisa Viso Gurovich por sus enseñanzas e ideas compartidas, pero especialmente, por renovar la admiración que desde estudiante me provocó su inteligencia y conocimientos. Por empujarme en aquellos años a descubrir mi pasión por la programación.

A mis sinodales por su tiempo y su experiencia.

A mi equipo de trabajo, por el privilegio que me brindan de coordinar a un grupo de excelencia. Personas sin las cuales, todas las ideas serían solo eso y no los hechos que han llegado a ser.

Al Instituto Nacional de Estadística y Geografía, por permitirme aportar a mi querido México un granito de arena para hacer mejor esta nación, con la esperanza de que algún día, el INEGI sea reconocido en este país como lo es fuera de él. Especial agradecimiento a los directivos que se la han jugado conmigo, confiando en mis propuestas.

A la Facultad de Ciencias de la UNAM y a mis profesores, por la formación académica que ha sustentado mi labor profesional.

A la Universidad Nacional Autónoma de México, que día a día me hace sentir orgullosamente universitario, de la UNAM.

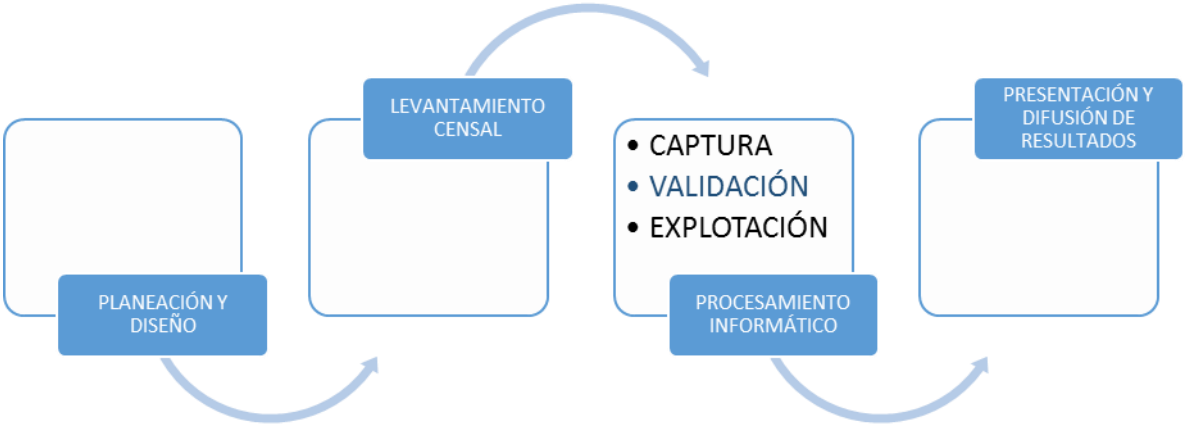
Contenido

Introducción

La realización de censos tiene una gran relevancia en el país pues mediante ellos, con una periodicidad determinada, se captan y posteriormente divulgan datos referentes a los aspectos más importantes de cada una de las unidades de observación de la población objeto de estudio. Esta información tiene un papel preponderante en la toma de decisiones en los aspectos económico, político y social.

El Instituto Nacional de Estadística y Geografía (INEGI) tiene la facultad exclusiva de realizar los Censos Nacionales[1] que se conforman en lo fundamental por el Conteo de Población y Vivienda, Censo General de Población y Vivienda, los Censos Económicos, Censos Agropecuarios y los Censos de Gobierno.

El proceso general de un censo consta de las siguientes etapas:



Una fase importante dentro de la etapa de Procesamiento Informático y en la que se circunscribe este trabajo, es la de Validación de la Información, proceso que tiene la finalidad de garantizar la integridad, congruencia y calidad de los datos captados respetando al máximo las respuestas del informante. Debido a la creciente demanda de información de calidad y oportunidad, se requiere que los datos captados en el levantamiento de los censos sean analizados en plazos cada vez más cortos, para lo cual se desarrollan aplicaciones informáticas *ad hoc* llamadas Sistemas de Validación. Dichos sistemas requieren para su construcción un conocimiento especializado en materia de sistemas de información, así como un buen entendimiento de los aspectos matemáticos implicados en el diseño conceptual del censo en cuestión. En la siguiente tabla se presentan los procesos principales de estos sistemas, entendiendo que algunos de ellos aplican únicamente para algún tipo específico de censo:

[1] Ley del Sistema Nacional de Información Estadística y Geográfica. Artículo 59.

Carga	Proceso en el que la información capturada se agrega en la base de datos diseñada para validar dicha información.
Filtro	En esta etapa se verifica que los datos de identificación cumplan con los criterios establecidos, así como que cada cuestionario contenga la información mínima necesaria.
Codificación Automática	Durante este proceso se convierten las respuestas de las preguntas abiertas del cuestionario a claves numéricas mediante un conjunto de procedimientos automáticos.
Codificación Manual	Para aquellos casos en los que no fue posible asignar la clave automáticamente por sistema, existe un módulo de codificación para la asignación manual de claves.
Clasificación	Para los Censos Económicos, se requiere este proceso a fin de ejecutar algoritmos para asignar el sector económico al que pertenece un establecimiento.
Normalización	Proceso de los Censos Agropecuarios cuya función es unificar a un sistema de medida común aquella información que haya sido registrada en unidades de medida diferente.
Consistencias Lógicas	Este proceso tiene como finalidad asegurar la consistencia y coherencia lógica de cada una de las respuestas y de la interrelación entre secciones del cuestionario, mediante la aplicación de los llamados criterios de validación.
Depurador Manual	Permite interactuar con la base de datos para introducir las adecuaciones pertinentes cuando un registro se reporta como inconsistente durante el proceso de evaluación de Consistencias Lógicas.
Depurador Masivo	Permite, a través de archivos, interactuar con la base de datos para realizar una enorme cantidad de adecuaciones cuando el Depurador Manual resulta insuficiente por la gran cantidad de cambios requeridos.
Reconsulta	Proceso en el que se realiza el intercambio de información entre analistas de oficinas centrales y analistas de la estructura territorial respecto a cuestionarios con inconsistencias.
Actualización Cartográfica	Proceso que garantiza una correcta referencia geográfica de la información estadística.

En particular y aunque los sistemas construidos bajo mi coordinación abarcan la totalidad de las etapas mencionadas, el presente trabajo se centra en el proceso de Consistencias Lógicas, mostrando su implementación en el Sistema de Validación de la Encuesta de Población y Vivienda 2015, destacando que en dicha implementación se ve reflejada la experiencia acumulada en el desarrollo de los Sistemas de Validación de los Censos de Población y Vivienda 2010, del Censo de Escuelas, Maestros y Alumnos de Educación Básica y Especial, así como en el de los Censos Económicos 2014.

Para garantizar la coherencia lógica de todas las respuestas a las preguntas del cuestionario de la Encuesta de Población y Vivienda 2015 así como de la interrelación entre éstas, fue utilizada la metodología de Vectores Teóricos que permite un control completo de los valores que pueden tomar un conjunto de variables, a través de la generación de todas las combinaciones posibles; adicionalmente es un lenguaje común entre los encargados del diseño conceptual de la validación y los desarrolladores de la aplicación informática.

Un vector teórico es un arreglo unidimensional cuyos componentes representan los valores que pueden tomar las variables del cuestionario; de esta forma, se generan tantos vectores como preguntas y relaciones entre ellas se desea validar. Las combinaciones de un vector teórico se obtienen combinando uno a uno los valores de cada uno de los componentes del vector hasta obtener el conjunto total de combinaciones. Una vez obtenido este conjunto, se crea una función de direccionamiento que permite asignar un valor a cada combinación con el fin de tomar una acción que puede ser la de reportar como inconsistencia de información, corrección automática o bien aceptar la combinación como un valor válido. Para desarrollar el sistema asociado, se diseñó un motor que evalúa las combinaciones y genera la función de direccionamiento, así como una arquitectura de software y hardware que permitió el decremento sustancial en los tiempos de procesamiento en la evaluación de los criterios de validación de la Encuesta de Población y Vivienda 2015[2]. Es precisamente éste el tema del presente trabajo en el que se detalla la plataforma, lenguaje, experiencias y las consideraciones adicionales para su implementación.

[2] En adelante, indistintamente se usará Encuesta de Población y Vivienda 2015 o EIC2015

Consistencias lógicas

Como se mencionó en la introducción, se desarrollan sistemas de validación en cada censo a fin de que los datos captados en la etapa levantamiento, sean analizados para garantizar su integridad y congruencia, mediante la aplicación de los llamados criterios de validación.

Fundamentalmente, los sistemas de validación deben proveer todos los mecanismos, funciones e interfaces para la revisión de las respuestas del cuestionario, con la esperanza de que contenga datos correctos, y de no ser así, permitir la imputación del valor adecuado de forma automática por el sistema (validación automática) o manual por un especialista, ambas formas usando ciertas reglas las cuales removerán las inconsistencias. Si no hay información suficiente y no es posible volver a consultar con el informante, los sistemas deben contar con el mecanismo de eliminación de registros que no pasaron alguna de las verificaciones o, al menos, omitir del análisis los campos que no fueron aprobados.

La validación automática se comprende como el conjunto de tratamientos que se aplica a la información proveniente de campo, y que tiene por objetivo eliminar las omisiones y respuestas inconsistentes en variables relacionadas entre sí. Para garantizar que para cada cuestionario de la Encuesta de Población y Vivienda 2015, los datos de identificación así como de referencia geográfica, cumplieran con los criterios establecidos, y que cada cuestionario contuviera la información mínima necesaria asegurando la relación lógica y aritmética entre sus variables y secciones relacionadas, fue utilizada la metodología de Vectores Teóricos, que es una herramienta que permite un control completo de los valores que pueden tomar un conjunto de variables, a través de la generación de todas las combinaciones posibles, la cual tiene en la exhaustividad una de sus principales características y ventajas, además de facilitar el control en la revisión de los resultados de la aplicación de los criterios de validación.

La experiencia en el análisis de la estadística que generan los sistemas de validación, es que ésta tiende a incorporar una cuantificación más precisa de los cambios que realiza, y busca una mayor desagregación y relación con el origen de los mismos, como producto de la aplicación de los criterios de validación en la información incorrecta a través de la metodología de vectores.

En lo que refiere a la Encuesta de Población y Vivienda 2015, se conservaron los elementos utilizados en censos anteriores, los que se emplearon para abordar los indicadores de análisis en la validación, a saber, la no respuesta, el cambio de código y el no especificado; utilizando como instrumentos a las matrices de entrada-salida, así como las frecuencias de imágenes y tratamientos. En seguida se presentan ejemplos de dichos instrumentos, formatos que fueron generados por el Sistema de Validación:

IDENTIFICACIÓN GEOGRÁFICA	
ENTIDAD:	01 - Aguascalientes
MUNICIPIO:	001 - Aguascalientes
NÚMERO DE MATRIZ:	11
VARIABLE:	DHSERSAL1
FECHA DE GENERACIÓN:	24/09/2015 05:36

MATRIZ ENTRADA - SALIDA
AFILIACIÓN A SERVICIOS DE SALUD (PRIMER OPCIÓN)
ABSOLUTOS

ENCUESTA INTERCENSAL 2015

9. ¿(NOMBRE) ESTÁ AFILIADA(O) O TIENE DERECHO A LOS SERVICIOS MÉDICOS EN: **(DHSERSAL1)**		TOTAL ENTRADA	SALIDA											
			BLANCO (b)	EL SEGURO POPULAR O PARA UNA NUEVA GENERACIÓN (Siglo XXI)? (1)	EL IMSS (Seguro Social)? (2)	EL ISSSTE? (3)	EL ISSSTE ESTATAL? (4)	PEMEX, DEFENSA O MARINA? (5)	UN SEGURO PRIVADO? (6)	OTRA INSTITUCIÓN ? (7)	ENTONCES, ¿NO ESTÁ AFILIADA(O) A SERVICIOS MÉDICOS? (8)	NO ESPECIFICADO (9)	TOTAL DE CAMBIOS	TOTAL
ENTRADA	BLANCO (b)	32	0	0	0	0	0	0	0	0	0	32	32	32
	EL SEGURO POPULAR O PARA UNA NUEVA GENERACIÓN (Siglo XXI)? (1)	12726	0	12726	0	0	0	0	0	0	0	0	0	12726
	EL IMSS (Seguro Social)? (2)	28827	0	0	28827	0	0	0	0	0	0	0	0	28827
	EL ISSSTE? (3)	3154	0	0	0	3154	0	0	0	0	0	0	0	3154
	EL ISSSTE ESTATAL? (4)	66	0	0	0	0	66	0	0	0	0	0	0	66
	PEMEX, DEFENSA O MARINA? (5)	88	0	0	0	0	0	88	0	0	0	0	0	88
	UN SEGURO PRIVADO? (6)	918	0	0	0	0	0	0	918	0	0	0	0	918
	OTRA INSTITUCIÓN? (7)	82	0	0	0	0	0	0	0	82	0	0	0	82
	ENTONCES, ¿NO ESTÁ AFILIADA(O) A SERVICIOS MÉDICOS? (8)	7345	0	1	0	1	0	0	0	0	7343	0	2	7345
	NO ESPECIFICADO (9)	0	0	0	0	0	0	0	0	0	0	0	0	0
	TOTAL	53238	0	12727	28827	3155	66	88	918	82	7343	32	34	53238

	REGISTRADO	TOLERANCIA	DIFERENCIA
OMISIÓN DE ENTRADA	0.06%	1.00%	-0.94%
OMISIÓN DE SALIDA	0.00%	0.00%	0.00%
NO ESPECIFICADO	0.06%	1.00%	-0.94%

ÍNDICE

SIN CAMBIO

FOCOS VS FOCOS AHORRADORES

ENCUESTA INTERCENSAL 2015

CRUCE DE VARIABLES
FOCOS / FOCOS_AHORRA

IDENTIFICACIÓN GEOGRÁFICA	
ENTIDAD:	01 - Aguascalientes
MUNICIPIO:	001 - Aguascalientes
NÚMERO DE CRUCE:	**61**
FECHA DE GENERACIÓN:	24/09/2015 05:31
VARIABLES:	**FOCOS / FOCOS_AHORRA**
REGISTROS PROCESADOS:	

Clase de vivienda particular (CLAVIVP) = 1, 2, 3, 4, 5, 6 y 99

11. ¿CUÁNTOS FOCOS TIENE ESTA VIVIENDA? **(FOCOS)**	11. ¿CUÁNTOS FOCOS SON AHORRADORES? **(FOCOS_AHORRA)**										
	BLANCO (b)	00	01	02	03	04	05 - 10	11 - 25	26 - 99	NO ESPECIFICADO (999)	TOTAL
BLANCO (b)	39	0	0	0	0	0	0	0	0	0	39
01	0	46	46	0	0	0	0	0	0	1	93
02	0	151	36	112	0	0	0	0	0	2	301
03	0	248	70	78	188	0	0	0	0	8	592
04	0	331	112	132	139	389	0	0	0	8	1111
05 - 10	0	1253	314	522	621	716	5311	0	0	35	8772
11 - 25	0	102	23	32	35	55	465	1825	0	4	2541
26 - 99	0	5	0	0	4	3	12	55	303	0	382
NO ESPECIFICADO (999)	0	1	0	0	0	0	0	0	0	19	20
TOTAL	39	2137	601	876	987	1163	5788	1880	303	77	13851

ÍNDICE

Es importante señalar que para este ejercicio estadístico fueron creados 562 reportes distintos, todos a nivel municipal, estatal y nacional en formato XLS (hojas de cálculo), lo que hace un total de 1,406,686 formatos generados en la etapa de producción del sistema.

METODOLOGÍA DE VECTORES TEÓRICOS

Definiciones

Consistencia lógica. Se refiere a la composición específica de variables y vectores con ciertos valores, mismos que interrelacionados conforman combinaciones. Cada una de estas combinaciones es asociada a una imagen, a la que a su vez se le asocia un tratamiento.

Tratamiento. Se define como un algoritmo en el cual se pueden asignar nuevos valores a una pregunta cuyo código de entrada se detecta como inconsistente.

Tratamientos por imagen. Se aplican exclusivamente a una combinación determinada de preguntas mediante un número de imagen ubicado por una función de direccionamiento, por lo que el tratamiento se aplica de forma específica a la imagen generada corrigiendo valores inconsistentes o señalando el registro como no consistente.

Vector teórico. Define las variables y condiciones que se manejarán al interior de estos arreglos.

Función de direccionamiento. Expresión matemática que se utiliza para obtener la referencia a la imagen que se está analizando para aplicarle un tratamiento específico.

Vectores teóricos

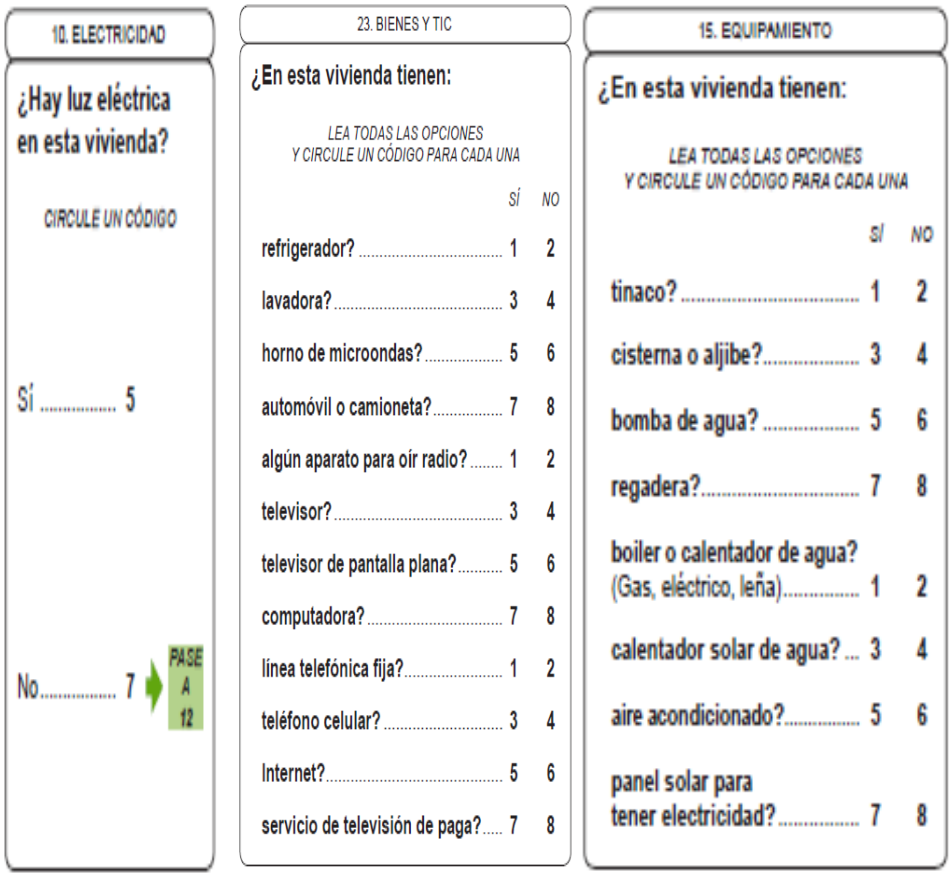
Un vector teórico V es un arreglo unidimensional cuyos componentes (V_1, V_2,...,V_n), representan los valores que pueden tomar las variables V_1, V_2, ...,V_n.

Ejemplo:

Se desea garantizar la consistencia entre ***Electricidad, Bomba de agua, Aire acondicionado, Refrigerador y Lavadora***. En caso de que la variable ***Electricidad*** tenga registrado el código 7 (No) o presente omisión, pero se tenga registrado que se dispone en la vivienda de al menos dos de los siguientes bienes ***Bomba de agua, Aire acondicionado, Refrigerador o Lavadora,*** entonces se asigna el código 5 (Sí) en la variable ***Electricidad***. Si se presenta omisión en la variable ***Electricidad*** y no se cuenta con información adicional que permita la asignación de algún código válido, se asigna el código 9 (No especificado) a esta variable.

CONDICIÓN DE ENTRADA**:** ***clase de vivienda particular*** ∈ {1,…, 6, 99} y ***electricidad*** ≠ 5

Variables en el cuestionario



10. ELECTRICIDAD

¿Hay luz eléctrica en esta vivienda?

CIRCULE UN CÓDIGO

Sí 5

No..... 7 → PASE A 12

23. BIENES Y TIC

¿En esta vivienda tienen:

LEA TODAS LAS OPCIONES Y CIRCULE UN CÓDIGO PARA CADA UNA

	SÍ	NO
refrigerador?	1	2
lavadora?	3	4
horno de microondas?	5	6
automóvil o camioneta?	7	8
algún aparato para oír radio?	1	2
televisor?	3	4
televisor de pantalla plana?	5	6
computadora?	7	8
línea telefónica fija?	1	2
teléfono celular?	3	4
Internet?	5	6
servicio de televisión de paga?	7	8

15. EQUIPAMIENTO

¿En esta vivienda tienen:

LEA TODAS LAS OPCIONES Y CIRCULE UN CÓDIGO PARA CADA UNA

	SÍ	NO
tinaco?	1	2
cisterna o aljibe?	3	4
bomba de agua?	5	6
regadera?	7	8
boiler o calentador de agua? (Gas, eléctrico, leña)	1	2
calentador solar de agua?	3	4
aire acondicionado?	5	6
panel solar para tener electricidad?	7	8

Variables y mnemónicos

VARIABLE	MNEMÓNICO
Electricidad	ELECTRICIDAD
Bomba de agua	BOMBA_AGUA
Aire acondicionado	AIRE_ACON
Refrigerador	REFRIGERADOR
Lavadora	LAVADORA

Entonces el vector V se define como $V = (V_1, V_2, V_3, V_4, V_5)$ donde

$V_1 =$ { 0 Si ELECTRICIDAD = ~~b~~[3]
1 Si ELECTRICIDAD = 7

$V_2 =$ { 0 Si BOMBA_AGUA ≠ 5
1 Si BOMBA_AGUA = 5

$V_3 =$ { 0 Si AIRE_ACON ≠ 5
1 Si AIRE_ACON = 5

$V_4 =$ { 0 Si REFRIGERADOR ≠ 1
1 Si REFRIGERADOR = 1

$V_5 =$ { 0 Si LAVADORA ≠ 3
1 Si LAVADORA = 3

Una vez definidos los vectores teóricos hay que determinar cuántas y cuáles son las combinaciones posibles.

Las combinaciones de un vector teórico se obtienen variando uno a uno los valores de cada uno de los componentes del vector, iniciando desde la última componente hasta la primera (de derecha a izquierda), y así obtener el conjunto total de combinaciones.

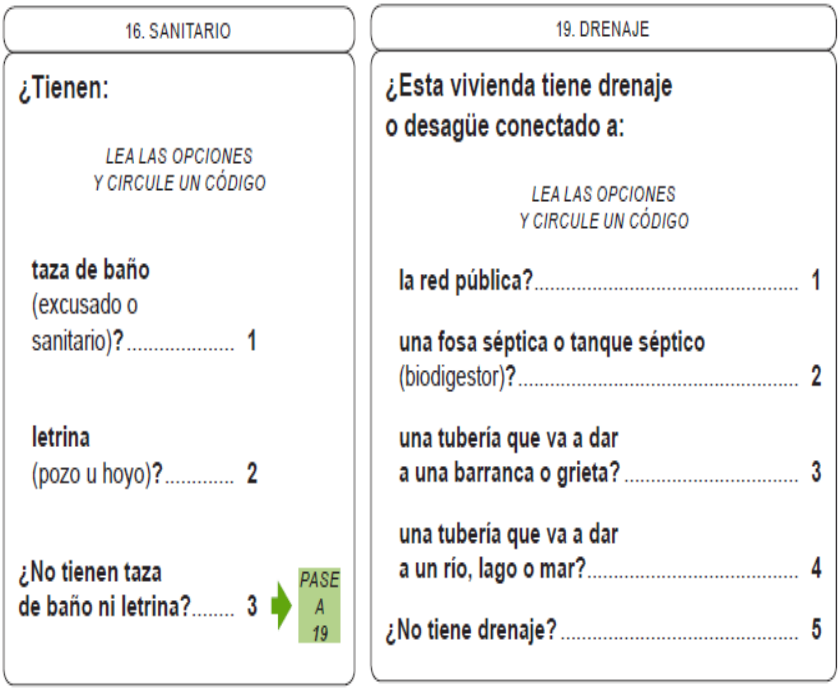
Ejemplo:

[3] ~~b~~ significa valor en blanco o ausencia de valor

Se desea garantizar la consistencia entre ***clase de vivienda particular***, ***servicio sanitario*** y ***drenaje***. Si en la vivienda particular con clave 3 (Casa dúplex, triple o cuádruple), 4 (Departamento en edificio) o 6 (Cuarto en la azotea de un edificio); se registra que tiene ***letrina*** o no tiene ***drenaje***, entonces se modifica la ***clase de vivienda particular*** a no especificada.

CONDICIÓN DE ENTRADA**: *clase de vivienda particular*** ∈ {3, 4, 6}

Variables en el cuestionario



Variables y mnemónicos

VARIABLE	MNEMÓNICO
Sanitario	SERSAN
Drenaje	DRENAJE

$V = (V_1, V_2)$ donde

$V_1 =$ { 0 Si SERSAN ≠ 2 ; 1 Si SERSAN = 2

$V_2 =$ { 0 Si DRENAJE ≠ 5 ; 1 Si DRENAJE = 5

Total de combinaciones 4, que son las siguientes:

(0 , 0) combinación 1

(0 , 1) combinación 2

(1 , 0) combinación 3

(1 , 1) combinación 4

O bien, obtenemos el total de combinaciones multiplicando entre sí el número de alternativas de cada vector teórico:

Para V_1 son dos (0 y 1) y para V_2 también 2, entonces, para el ejemplo, el total de combinaciones es $2 \cdot 2 = 4$.

A fin de verificar el orden de las combinaciones y facilitar su manejo, se crea la función de direccionamiento, la cual permite asignar un valor a cada combinación y que corresponde al orden en que ésta es generada. Su fórmula general es:

$$f(V_1,V_2,\ldots,V_n) = C_1 \cdot V_1 + C_2 \cdot V_2 + \ldots + C_n \cdot V_n + 1 = \sum C_i \cdot V_i + 1 \text{ para } i=1..n$$

Donde

V_i son los componentes del vector teórico.

C_i son los coeficientes de la función de direccionamiento, calculados con el siguiente algoritmo:

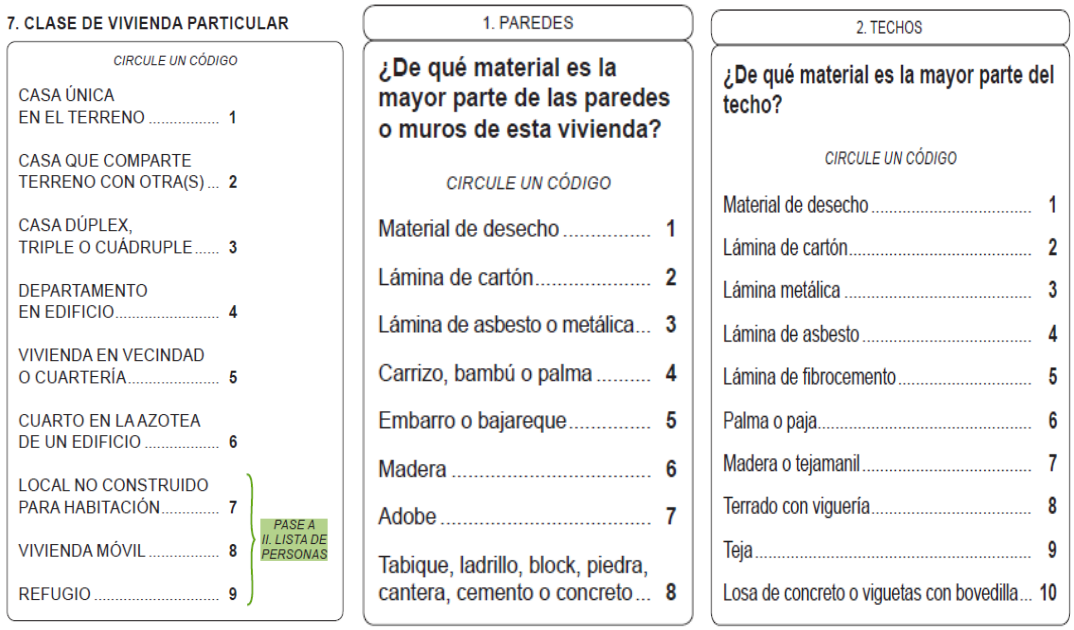
1. Dividir el número total de combinaciones entre el número de alternativas del primer componente del vector (V_1). El resultado o cociente obtenido, será el coeficiente (C_1) del primer vector.
2. Dividir el coeficiente C_k entre el número de alternativas del componente k+1 del vector (V_{k+1}). El resultado o cociente obtenido será el coeficiente (C_{k+1}), para k = 1..n-1 o hasta que el coeficiente obtenido sea 1, que corresponderá al coeficiente del último componente del vector.

Ejemplo:

Se desea garantizar la congruencia entre las variables ***Clase de Vivienda Particular, Paredes y Techos***. Cuando la variable ***Clase de Vivienda Particular***, es del tipo departamento en edificio y la respuesta en las variables ***Paredes*** y ***Techos*** corresponde a materiales precarios, se les debe asignar a estas últimas el código 9 (No especificado) y el código 99 (No especificado), respectivamente. En caso de omisión en las variables ***Paredes*** y ***Techos***, se les debe asignar el código 9 (No especificado) y el código 99 (No especificado), respectivamente.

CONDICIÓN DE ENTRADA**: *clase de vivienda particular*** ∈ {1,..., 6, 99}

Variables en el cuestionario



7. CLASE DE VIVIENDA PARTICULAR

CIRCULE UN CÓDIGO

CASA ÚNICA EN EL TERRENO 1
CASA QUE COMPARTE TERRENO CON OTRA(S) ... 2
CASA DÚPLEX, TRIPLE O CUÁDRUPLE 3
DEPARTAMENTO EN EDIFICIO 4
VIVIENDA EN VECINDAD O CUARTERÍA 5
CUARTO EN LA AZOTEA DE UN EDIFICIO 6
LOCAL NO CONSTRUIDO PARA HABITACIÓN 7
VIVIENDA MÓVIL 8
REFUGIO 9

PASE A II. LISTA DE PERSONAS

1. PAREDES

¿De qué material es la mayor parte de las paredes o muros de esta vivienda?

CIRCULE UN CÓDIGO

Material de desecho 1
Lámina de cartón 2
Lámina de asbesto o metálica ... 3
Carrizo, bambú o palma 4
Embarro o bajareque 5
Madera 6
Adobe 7
Tabique, ladrillo, block, piedra, cantera, cemento o concreto ... 8

2. TECHOS

¿De qué material es la mayor parte del techo?

CIRCULE UN CÓDIGO

Material de desecho 1
Lámina de cartón 2
Lámina metálica 3
Lámina de asbesto 4
Lámina de fibrocemento 5
Palma o paja 6
Madera o tejamanil 7
Terrado con viguería 8
Teja 9
Losa de concreto o viguetas con bovedilla ... 10

Variables y mnemónicos

VARIABLE	MNEMÓNICO
Clase de Vivienda Particular	CLAVIVP
Paredes	PAREDES
Techos	TECHOS

$V = (V_1, V_2, V_3)$ donde

$V_1 =$ { 0 Si CLAVIVP ∈ {1, 2, 5, 6, 99}; 1 Si CLAVIVP ∈ {3, 4} }

$V_2 =$ { 0 Si PAREDES = ~~b~~; 1 Si PAREDES ∈ {1,...,7}; 2 Si PAREDES = 8 }

$V_3 =$ { 0 Si TECHOS = ~~b~~; 1 Si TECHOS ∈ {1,...,9}; 2 Si TECHOS = 10 }

Siguiendo el algoritmo para la generación de la función de direccionamiento:

1- Total de combinaciones = $2 \cdot 3 \cdot 3 = 18$

2- $C_1 = 18 / 2 = 9$
$C_2 = 9 / 3 = 3$
$C_3 = 3 / 3 = 1$

Por lo tanto $f(V_1,V_2,V_3) = 9V_1 + 3V_2 + V_3 + 1$

Evaluando la función f, arroja las 18 imágenes que representan la totalidad de combinaciones:

f(0,0,0) = 1	f(0,0,1) = 2	f(0,0,2) = 3
f(0,1,0) = 4	f(0,1,1) = 5	f(0,1,2) = 6
f(0,2,0) = 7	f(0,2,1) = 8	f(0,2,2) = 9
f(1,0,0) = 10	f(1,0,1) = 11	f(1,0,2) = 12
f(1,1,0) = 13	f(1,1,1) = 14	f(1,1,2) = 15
f(1,2,0) = 16	f(1,2,1) = 17	f(1,2,2) = 18

Una vez que se ha calculado la función de direccionamiento, lo siguiente es analizar las combinaciones resultantes a fin de definir un cuadro de imágenes y procedimientos que especificará, para cada una de las combinaciones posibles, la aceptación, corrección automática o envío a depuración manual de información. Para ello, se siguen los pasos descritos a continuación:

1. Analizar cada una de las combinaciones obtenidas, las cuales nos dan todos los casos posibles de información.

2. Definir criterios de validación para determinar:

    - Las combinaciones de información correcta.
    - Las combinaciones de información inconsistente que pueden ser corregidas automáticamente.
    - Los criterios de corrección automática.
    - Las combinaciones con información inconsistente que deberán ser enviadas a depuración manual.

3. Agrupar imágenes en inconsistentes para envío a depuración manual, correctas, e inconsistentes para corrección automática; para estas últimas, establecer como segundo criterio de agrupación su procedimiento de corrección.

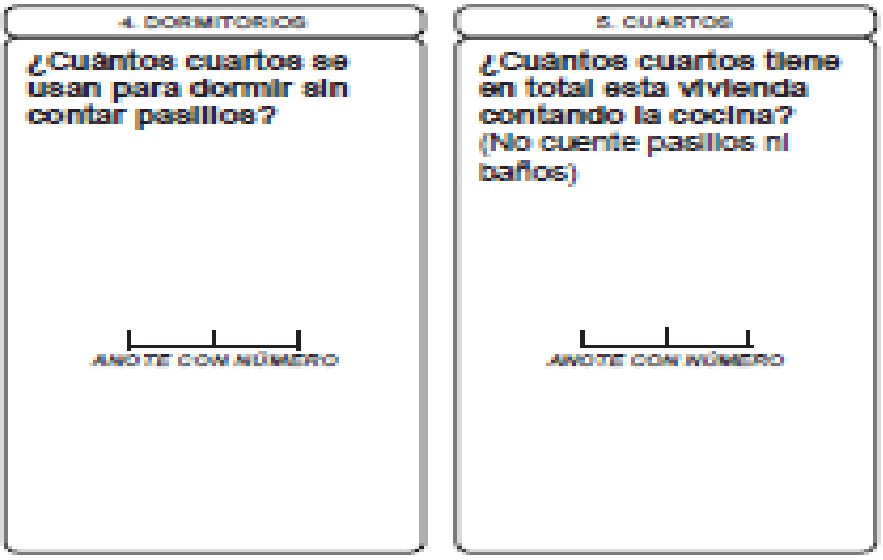
4. Generar un cuadro, con al menos dos columnas, en el cual se especificarán las imágenes y en su fila correspondiente el criterio de validación (procedimiento) definido para su tratamiento.

Ejemplo:

Se desea garantizar la consistencia entre los ***Dormitorios*** y ***Cuartos***. Si la variable cuartos toma los valores 1 o 2 y la variable ***Dormitorios*** presenta información incongruente, se asigna a esta última el valor 1. Cuando las variables ***Cuartos*** o ***Dormitorios*** presentan valores mayores al máximo permitido (25), omisión o sean iguales a 0 (cero), se les asigna el código 99 (No especificado).

CONDICIÓN DE ENTRADA**: *clase de vivienda particular*** $\in$ {1,..., 6, 99}

Variables en el cuestionario



Variables y mnemónicos

VARIABLE	MNEMÓNICO
Dormitorios	CUADORM
Cuartos	TOTCUART

$V = (V_1, V_2)$ donde

$V_1 =$

0 Si CUADORM ∈ {~~b~~, 0}

1 Si CUADORM ∈ {26,…,99}

2 Si CUADORM ∈ {1,…,25}

$V_2 =$

0 Si TOTCUART ∈ {~~b~~, 0}

1 Si TOTCUART ∈ {26,…,99}

2 Si TOTCUART ∈ {1, 2}

3 Si TOTCUART ∈ {3,…,25}

Función de direccionamiento

$F(V_1, V_2) = 4\,V_1 + V_2 + 1$

Imágenes para la combinación de variables y procedimientos

NO. IMAGEN	CUADORM	TOTCUART	PROCEDIMIENTO
1	∈ {~~b~~, 0}	∈ {~~b~~, 0}	Asignar CUADORM = 99 y TOTCUART = 99
2	∈ {~~b~~, 0}	∈ {26,...,99}	Asignar CUADORM = 99 y TOTCUART = 99
3	∈ {~~b~~, 0}	∈ {1, 2}	Asignar CUADORM = 1
4	∈ {~~b~~, 0}	∈ {3,...,25}	Asignar CUADORM = 99
5	∈ {26,...,99}	∈ {~~b~~, 0}	Asignar CUADORM = 99 y TOTCUART = 99
6	∈ {26,...,99}	∈ {26,...,99}	Asignar CUADORM = 99 y TOTCUART = 99
7	∈ {26,...,99}	∈ {1, 2}	Asignar CUADORM = 1
8	∈ {26,...,99}	∈ {3,...,25}	Asignar CUADORM = 99
9	∈ {1,...,25}	∈ {~~b~~, 0}	Asignar TOTCUART = 99
10	∈ {1,...,25}	∈ {26,...,99}	Asignar TOTCUART = 99
11	∈ {1,...,25}	∈ {1, 2}	Sin cambio
12	∈ {1,...,25}	∈ {3,...,25}	Sin cambio

Realizando la agrupación de imágenes tenemos:

NO. IMAGEN	PROCEDIMIENTO
11, 12	Sin cambio
1, 2, 5, 6	Asignar CUADORM = 99 y TOTCUART = 99
4, 8	Asignar CUADORM = 99
9, 10	Asignar TOTCUART = 99
3, 7	Asignar CUADORM = 1

La utilización de la metodología de Vectores Teóricos para las Consistencias Lógicas de los Sistemas de Validación de Información Censal, presenta las siguientes ventajas:

1.- Se consideran todos y cada uno de los casos posibles.

2.- Para cada caso se asocia un único tratamiento.

3.- Da mayor claridad en la reducción de procesos repetidos.

4.- Da mayor claridad en la identificación de problemáticas y en consecuencia, en la resolución de los mismos.

5.- La experiencia obtenida en su utilización nos ha dado mayor agilidad en la elaboración de los procesos de validación.

Implementación de la evaluación de Criterios de Validación

Una vez que las áreas conceptuales decidieron usar la metodología de Vectores Teóricos para la validación de la información de la Encuesta de Población y Vivienda 2015, el reto siguiente para la parte informática fue su implementación en un sistema que permitiera ejecutar el proceso para los datos captados: más de 5.8 millones de viviendas y arriba de 22.6 millones de personas, así como la generación del gran número de reportes derivados de ésta en un máximo de 48 horas.

Para poder superar tan enorme desafío, el grupo de desarrollo cumplimos fundamentalmente con lo siguiente:

- La materialización de nuestra experiencia en la construcción de sistemas de validación utilizando la metodología de Vectores Teóricos, en los que se han analizado y, en su momento, programados y ejecutados más de 1,500 criterios de requerimientos de los distintos censos (de población, económicos, agropecuarios y educativos), que han tenido lugar a partir de 2004.

- El diseño de una arquitectura abierta (lenguaje, estructuras, plataformas y recursos de cómputo en general), apegada a las normas institucionales, que permitió articular los componentes que fueron trabajados por separado en grupos de desarrollo geográficamente dispersos, en Aguascalientes, Puebla y DF, y que aseguró contar con bases de datos en repositorios únicos y centralizados.

- El diseño arquitectónico de la aplicación, que permitió una mejor distribución del trabajo de procesamiento y uso de recursos en general, con la capacidad para utilizar servidores productivos virtuales que se adicionaran posteriormente. Esto significa que a través de parámetros fue posible acceder a diferentes instancias de bases de datos y contenedores de aplicaciones, según la demanda de recursos dependiendo de las necesidades ya en la etapa productiva del sistema.

- Asociado a las consistencias lógicas, el diseño y programación de un motor que evalúa las combinaciones y genera la función de direccionamiento de forma automática.

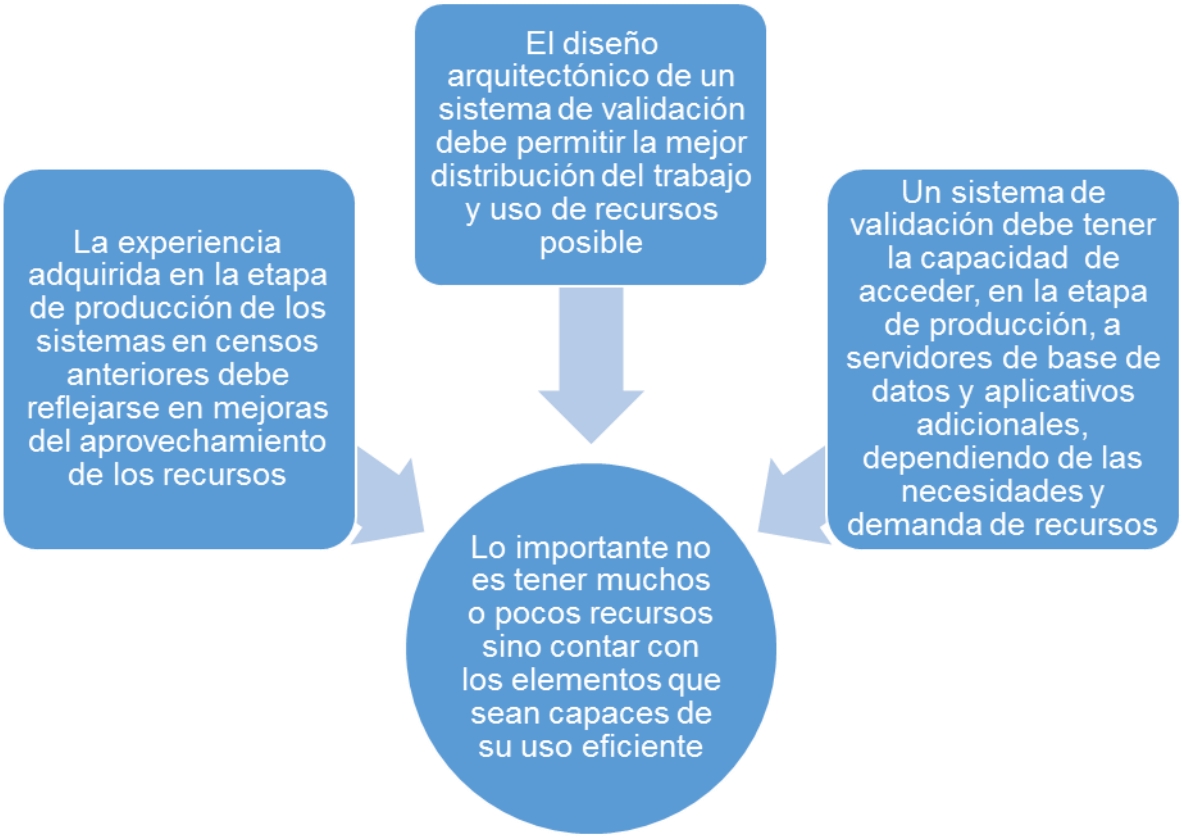
- Dotar al sistema con la capacidad para ejecutar una unidad de procesamiento adecuada a cada situación, dependiendo de las necesidades del momento y en función de la información; esto es, poder procesar cualquier unidad derivada de una clasificación lógica de los datos, por ejemplo por lote, entidad, municipio, localidad, tipo de vivienda, etc., o incluso un cuestionario en específico.

- Para el mejor aprovechamiento del proceso en paralelo, el manejo parametrizado del número de hilos (*threads*) que se ejecutaron en cada secuencia, dependiendo de las necesidades y recursos con los que se contó en un momento dado, pudiendo ensayar distintos escenarios dependiendo de la cantidad de memoria disponible, o sea mayor o menor número de cuestionarios procesándose al mismo tiempo.

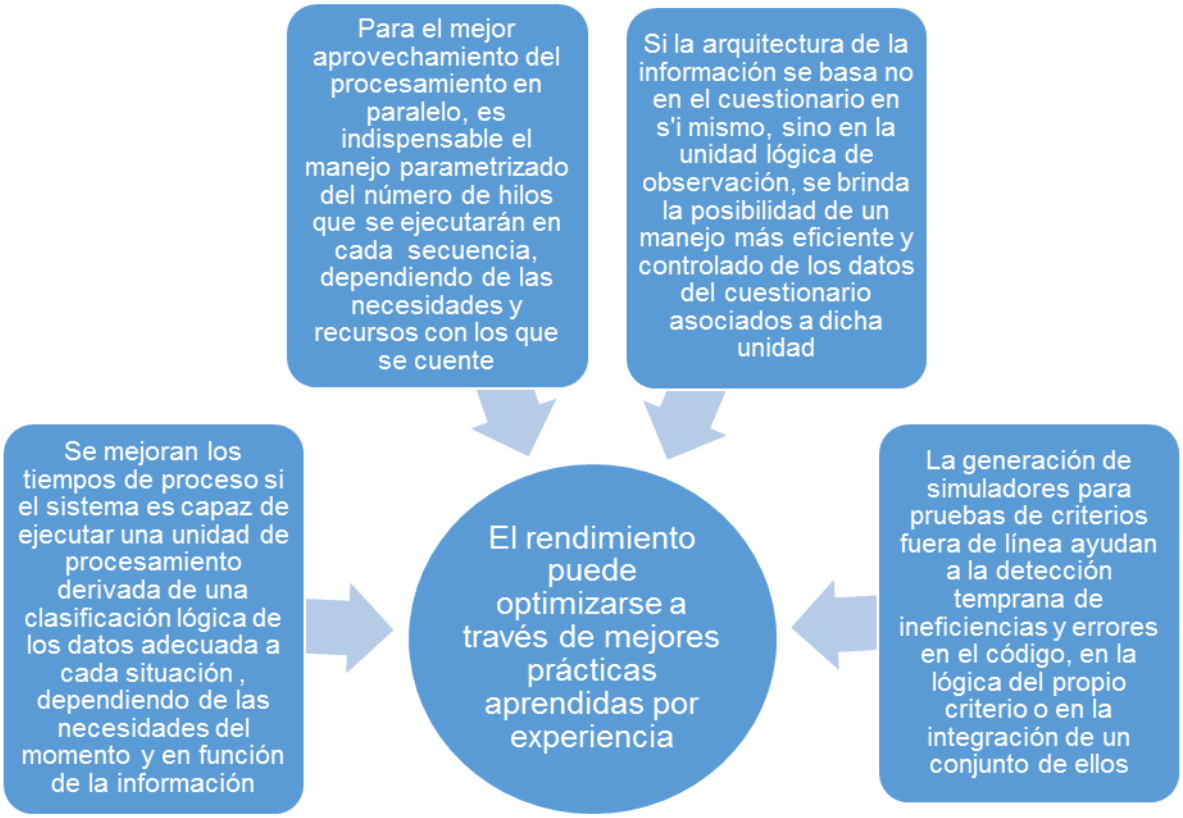
- La generación de un simulador para pruebas de evaluación de criterios fuera de línea, que ayudaron a la detección temprana de errores en el código o en la lógica del propio criterio, o integración de un conjunto de ellos. Sustituyendo además los largos y complicados períodos de pruebas, que comprendían una larga etapa de capacitación a grupos especialmente dedicados a ello y que, con los simuladores, dichas pruebas las realizaron directamente el personal que diseñó los criterios.

- Una apropiada puesta a punto de los servidores a través de los parámetros adecuados para este censo en especial, considerando el tamaño y número de registros a validar.

Consideraciones de uso de recursos



Consideraciones de optimización



En lo que respecta a la arquitectura de la aplicación, que fue multicapa, se separó la lógica de la funcionalidad de la lógica del diseño, específicamente utilizando como modelo de abstracción de desarrollo de software el patrón Modelo Vista Controlador (MVC), a fin de separar los datos, las interfaces y la lógica de negocios. El Modelo es el sistema de gestión de base de datos así como la lógica de negocio; la Vista son las interfaces de usuario; y el Controlador es el responsable de recibir los eventos de entrada desde la vista, invocando peticiones al Modelo y a la Vista. En particular, para la funcionalidad de la validación automática de la EIC2015, al ser un tema de procesamiento masivo de datos, sólo requirió una interfaz que presentaba la bitácora de procesamiento, acompañando a la opción de lanzamiento del proceso. De tal forma que a través de ésta interfaz, fue posible monitorear el avance en la corrida de la validación, toda vez que en ella se iba mostrando en tiempo real el número de viviendas validadas.

El proyecto fue programado en lenguaje Java así como en PL/SQL (lenguaje nativo en la base datos), y los datos fueron alojados en el Sistema Gestor de Bases de Datos Relacionales Oracle, respetando las siguientes tres capas:

Capa de presentación

Para crear las interfaces gráficas se utilizó el *framework*[4] *ZK* versiones *5.1 y 6,* generando contenido dinámico para web mediante *JSP*[5]*.*

Capa de negocio

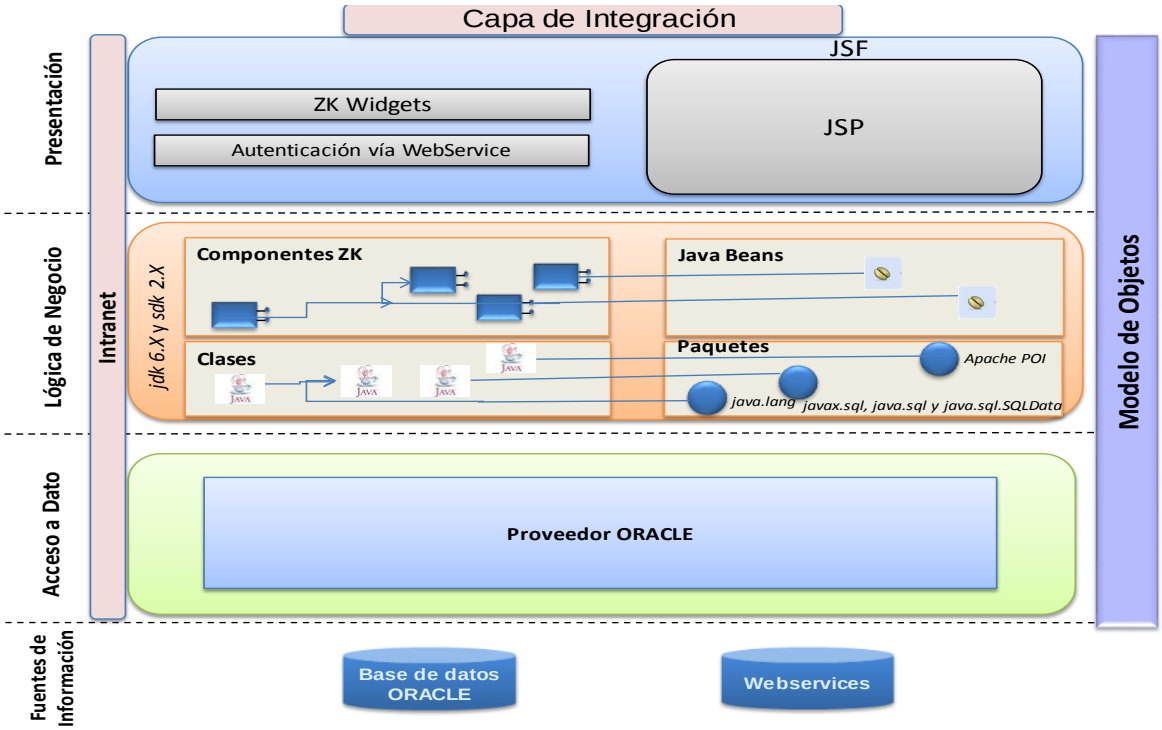
Para la creación de programas se recurrió al conjunto de herramientas (programas y bibliotecas) jdk 6.X y sdk 2.X.

Como API (Application Programming Interface), biblioteca Java para interactuar con documentos (hojas de cálculo) de Microsoft, se usó Apache POI.

Para lo referente al acceso de datos, fueron utilizados esencialmente los paquetes javax.sql, java.sql y java.sql.SQLData.

Capa de datos

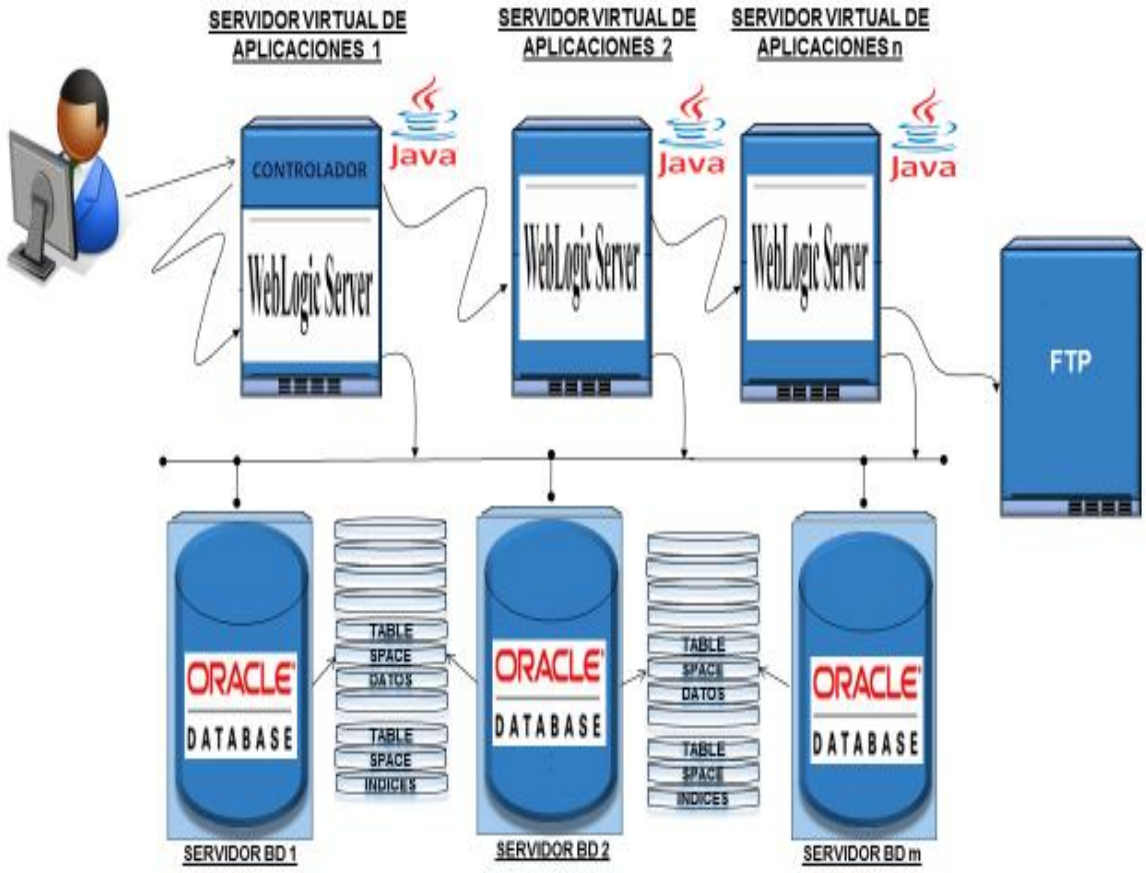
Fue empleado el servidor de base de datos Oracle 11g.



[4] El término se refiere a una estructura de software compuesta de componentes personalizables e intercambiables para el desarrollo de una aplicación.

[5] Es un acrónimo de Java Server Pages, tecnología orientada a crear páginas web dinámicas basadas en HTML, XML, entre otros tipos de documentos con programación en Java.

La siguiente lámina presenta la arquitectura genérica diseñada para el Sistema de Validación de la EIC2015.



La idea central fue la de tener la flexibilidad de usar, dependiendo de las necesidades de procesamiento, los recursos propicios para cubrirlas. A través de parámetros, gestionar la definición de servidores virtuales a acceder. De la misma forma para el tema de las bases de datos, en caso de ser necesario, realizar alguna partición lógica de información, a fin de alojar las distintas divisiones en servidores separados, con sus propios recursos, para mejorar ostensiblemente el rendimiento.

En el caso particular de la evaluación de los criterios de validación de la EIC2015, se utilizaron 2 servidores de aplicaciones virtuales, uno con 12 GB y otro con 5 GB de memoria RAM, así como un servidor de base de datos con 20 GB de memoria RAM y 160 GB de espacio en discos. También fue utilizado un servidor FTP, que alojó los archivos de los reportes, con el objeto de dejarlos disponibles al usuario[6].

[6] Requiriente del desarrollo del sistema y experto en los temas conceptuales de la Encuesta de Población y Vivienda 2015.

Mi experiencia en el desarrollo de sistemas de validación me indica que dadas las características siempre distintas de cada censo (aunque tienen procesos comunes, son diferentes entre sí), la complejidad de la mayoría de los criterios de validación y los tiempos requeridos de ejecución cada vez más cortos, la mejor opción no es desarrollar **un** sistema general de validación, sino uno que genere el código específico para cada censo, siempre considerando que la arquitectura, así como la puesta a punto de servidores (también distintos en cada ocasión), debe ser la adecuada, específica para cada censo. De otra forma, la solución tecnológica no respondería exclusivamente a las necesidades y requerimientos de las áreas conceptuales que, en la etapa de planeación y diseño, reflejan el dinamismo del país; por el contrario, dichas áreas tendrían que adecuarse a la forma de operar del sistema general, lo cual es inaceptable.

Tomando en cuenta lo anterior, para la EIC2015 se desarrollaron generadores de código, que son la propuesta de solución tecnológica para la construcción de sistemas de validación a la medida de cada censo, basada en el re uso de componentes. A través de ellos, se logró la creación automática del Simulador, cuyos criterios al ser liberados por el usuario fueron insertados sin cambio alguno en el Sistema de Validación en producción.

Los generadores contemplaron la producción de código de:

- Creación de los elementos en la base de datos: tablas, tipos, objetos y procedimientos almacenados.
- Mecanismos de intercambio de datos entre el servidor de base de datos y el servidor de aplicaciones.
- Mapeadores, que son piezas de software que permiten asignar información de una fuente origen a una fuente destino, empatando los tipos y tamaños de datos. Estos artefactos entregan datos solicitados por el destino, trayéndolos desde el origen.
- Estructura de los módulos para procesamiento paralelo (multiproceso).

Con respecto a la persistencia de la información, fueron considerados los siguientes aspectos importantes:

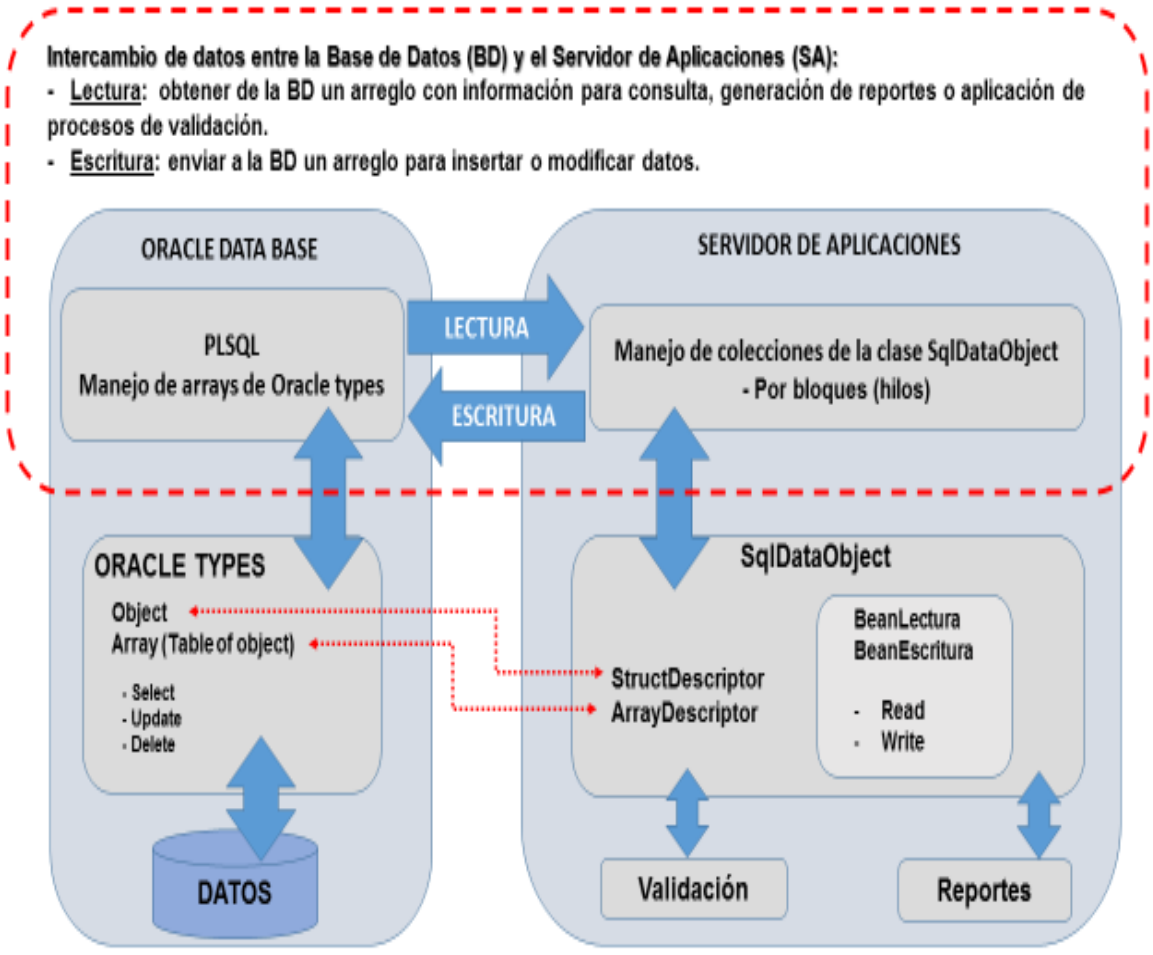
- Los usuarios deben poder acceder a los datos, en cualquier etapa del proyecto, de diferentes maneras y con diferentes herramientas.
- El almacenamiento de la información requiere tener un altísimo desempeño y confiabilidad. Alta velocidad de lectura y escritura.
- El desarrollo fue en Java, que es un lenguaje orientado a objetos, y el desarrollo de las bases de datos relacionales (RDBMS) es previo al paradigma de objetos, por lo que consecuentemente el soporte y

rendimiento en estos motores es, en general, bastante pobre en relación al paradigma citado.

- El mapeo entre objetos y relaciones no es un asunto trivial dadas las diferencias entre ambos modelos. Fundamentalmente, la teoría relacional se centra en los datos mientras que la de objetos se enfoca en el comportamiento.
- La evaluación de la utilización del *framework* de mapeo objeto-relacional más usado en Java, Hibernate, que resultó lento para las necesidades de la aplicación, pues fue construido básicamente para sistemas de interacción con el usuario y no para procesamiento masivo de datos.

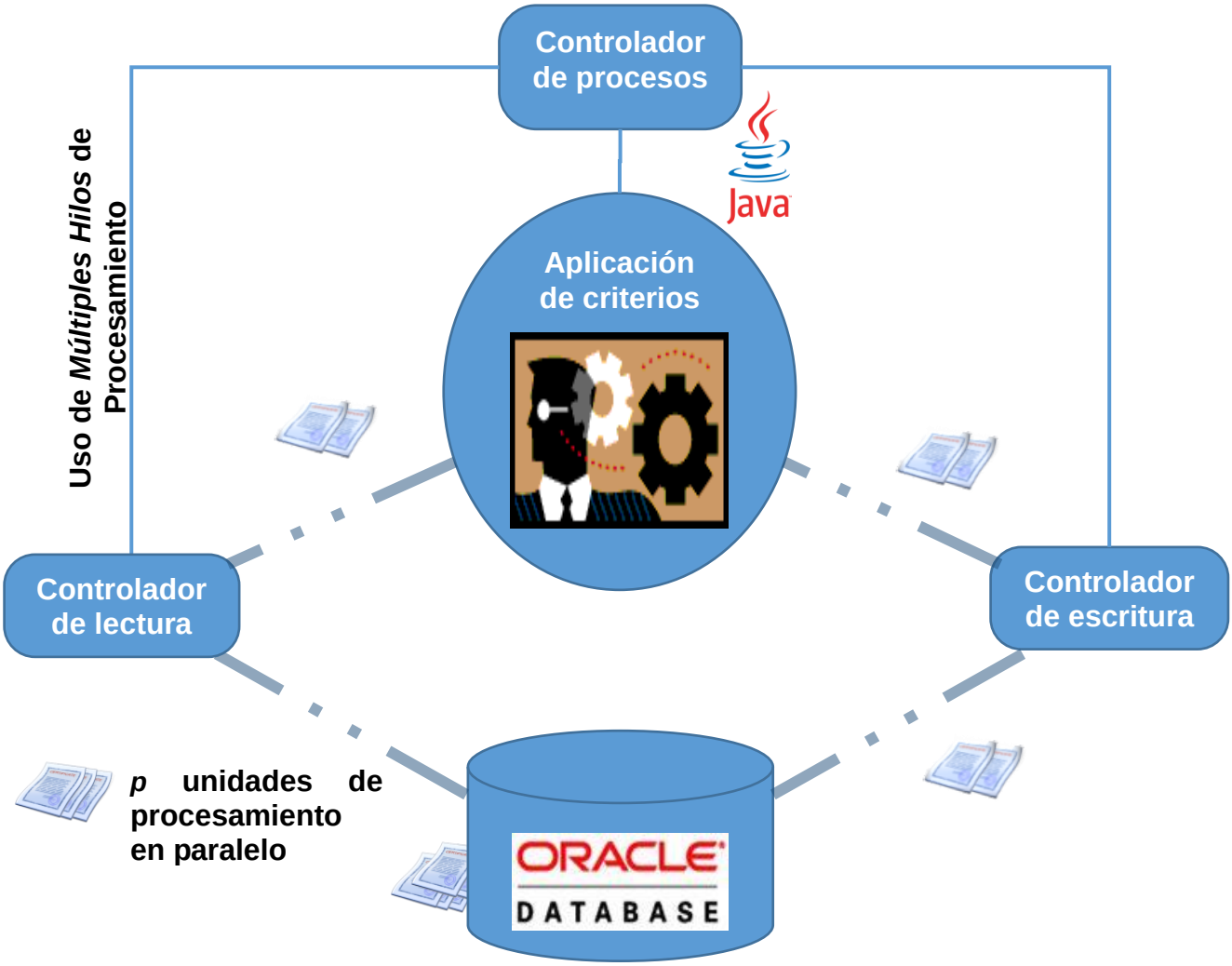
Dadas las consideraciones anteriores, la decisión fue construir nosotros mismos el mecanismo de intercambio de datos entre la aplicación y la base de datos, basados en el siguiente esquema:

OBJETOS ORACLE - JAVA



Para implementar las consistencias lógicas a través de vectores teóricos, una de las primeras acciones fue verificar el adecuado diseño de las estructuras de datos, tanto para la aplicación Java como para la capa de persistencia. A fin de hacer eficiente el traslado de información entre ambas, se diseñaron objetos similares para las dos capas y los artefactos de software que los soportan. De esta forma en la memoria del servidor de aplicaciones tenemos una imagen de los tipos de datos objeto de Oracle. Esto es, tanto en la aplicación como en la base de datos, tenemos la representación de unidades lógicas de observación, o sea, viviendas con sus personas asociadas, a las cuales se les van a aplicar los procesos de validación. Esto nos permite leer y escribir de y a la base de datos colecciones completas de dichas unidades, a través del intercambio de arreglos de los objetos que los representan. Adicionalmente, nos permite la utilización de multiproceso al tener unidades completas en la memoria; así, en un momento dado, se le aplican las consistencias lógicas en paralelo a varias de ellas.

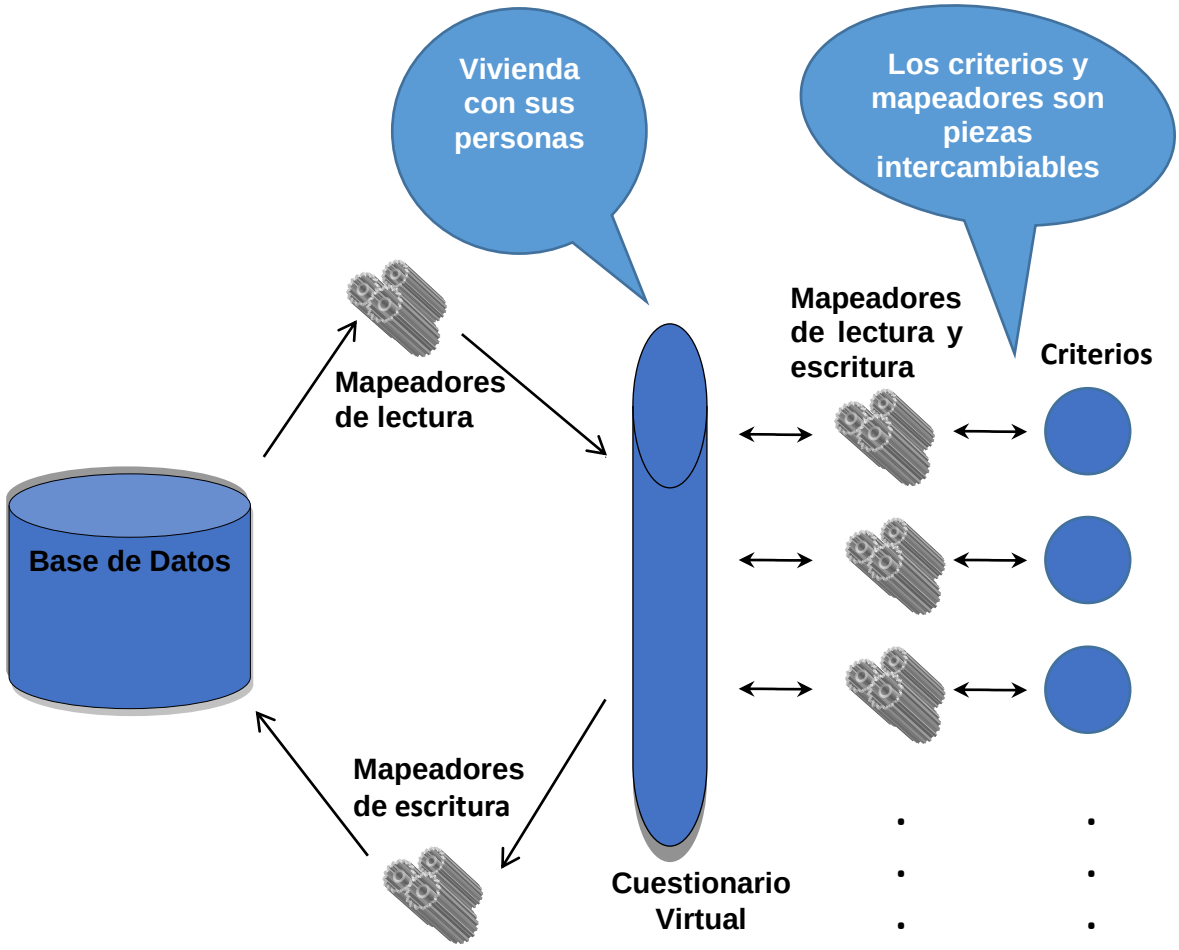
PROCESAMIENTO EN PARALELO



En la imagen que se presenta a continuación podemos observar el diseño conceptual para la evaluación de criterios de validación. Se tiene un origen de la información o fuente, la base de datos, que a través de mapeadores de lectura proveen información al Cuestionario Virtual, que es la representación de una vivienda con sus personas, conteniendo los datos que el informante proporcionó al respecto. A través de mapeadores de lectura y escritura, cada criterio recibe los campos que requiere, tanto para la evaluación que ejecuta, como para el resultado con el que se actualiza el Cuestionario Virtual, que a su vez, con mapeadores de escritura, reemplaza valores en la base de datos (destino) con atributos o campos actualizados por el proceso de la evaluación de los criterios de validación. De esta forma, la base de datos es fuente y destino, dependiendo del momento en que el proceso se encuentre.

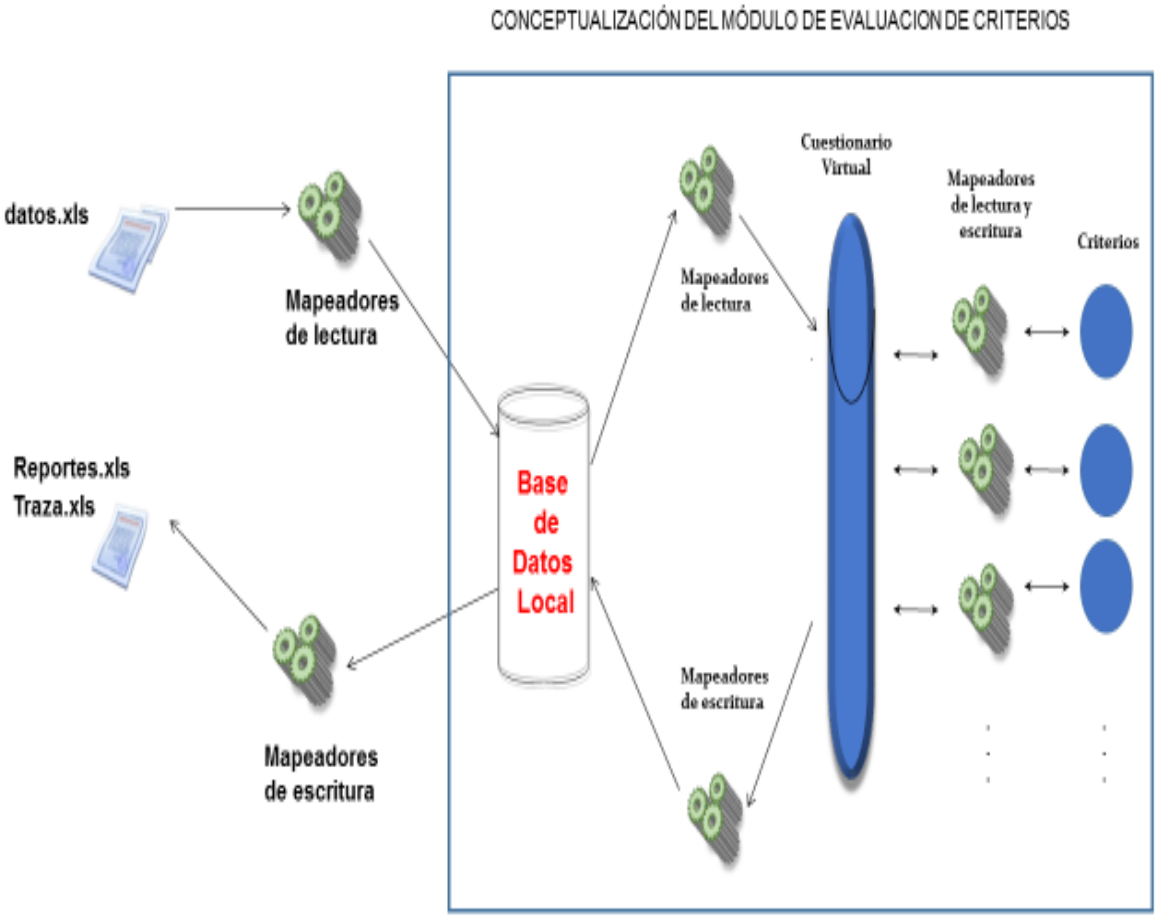
Es importante destacar que tanto los mapeadores como los criterios son piezas intercambiables, lo que permite el desarrollo por separado de cada uno de ellos. Esto da flexibilidad en la construcción del sistema en su conjunto al permitir añadir, sustituir o eliminar dichas piezas (que son clases en java) y permite la fácil realización de pruebas, tanto conceptuales como de funcionamiento.

CONCEPTUALIZACIÓN DEL MÓDULO DE EVALUACION DE CRITERIOS



Este diseño facilitó, además, la generación de un simulador de pruebas de evaluación de criterios fuera de línea, para la detección de errores tanto en la programación como en la lógica de los mismos criterios. Permitió la posibilidad de la realización de las pruebas por parte del mismo personal que diseñó los vectores teóricos.

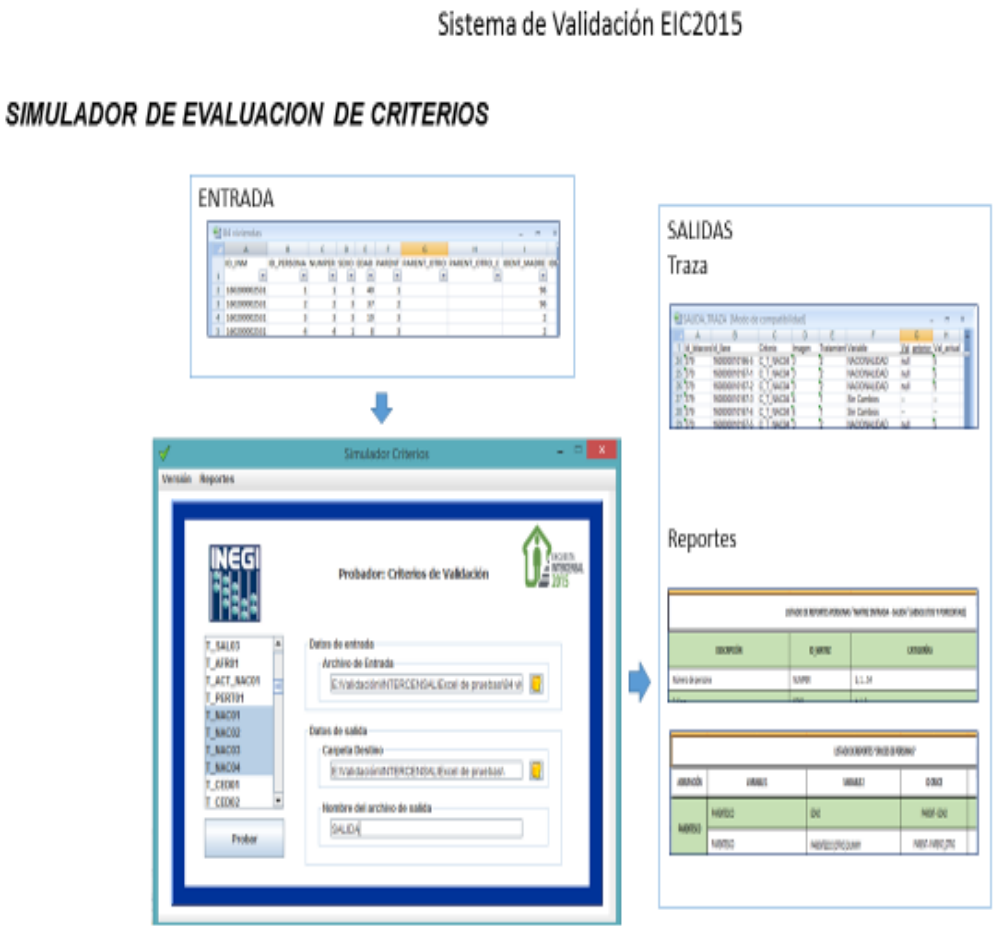
CONCEPTUALIZACIÓN DEL SIMULADOR DE EVALUACION DE CRITERIOS



Como se puede notar, se añadió la parte izquierda, que fueron mapeadores de lectura y escritura de y hacia hojas de cálculo, mismos que fueron creados por el usuario, con datos especialmente generados para que el simulador instalado en su equipo personal le permitiera verificar el comportamiento de los criterios diseñados por él mismo. Nótese que la fuente y destino, en este caso, se convirtió en una base de datos local, pero completamente idéntica a la del servidor institucional, únicamente con menor capacidad de almacenamiento y procesamiento, sólo para realizar pruebas, no el proceso completo.

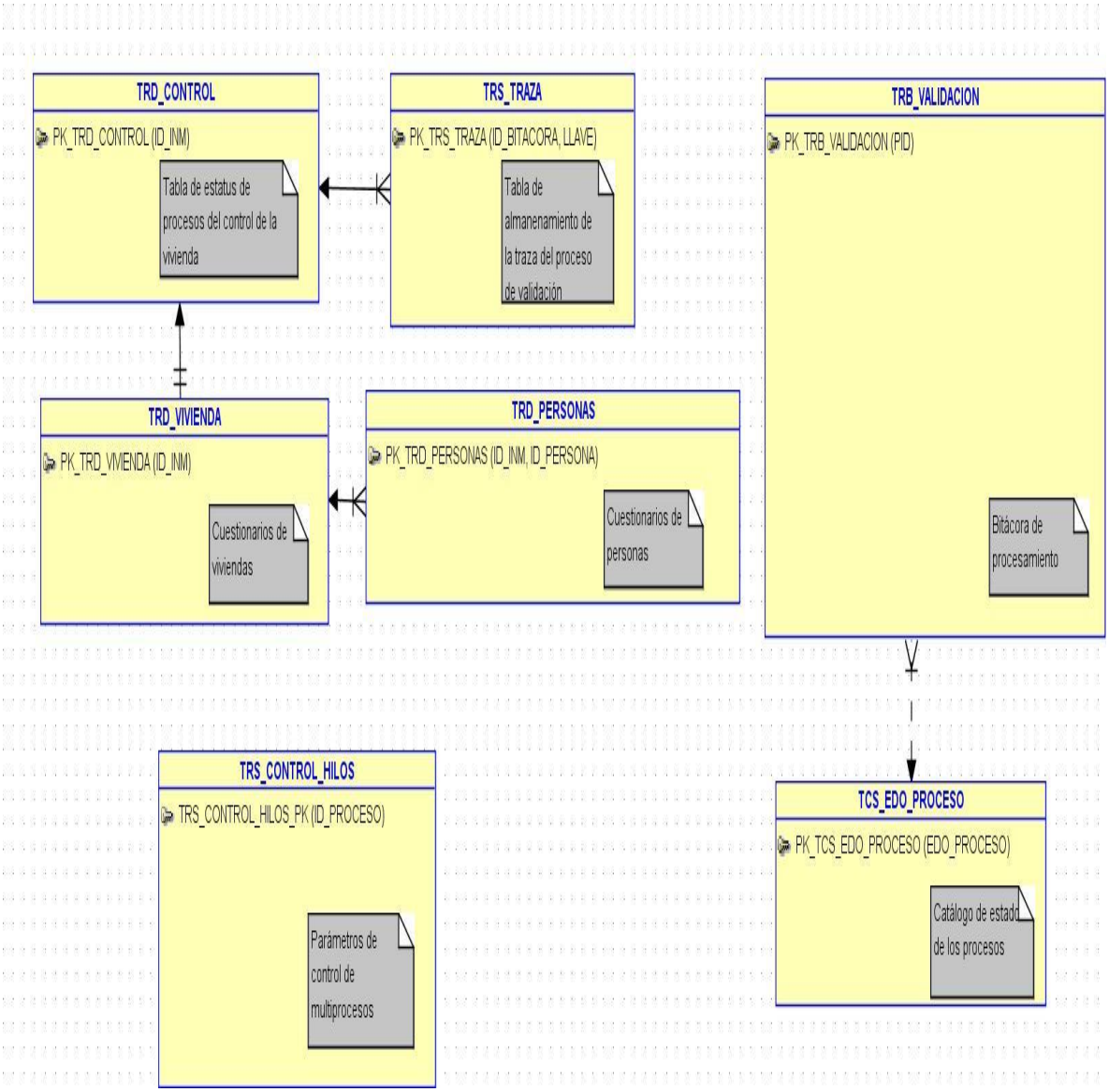
Las salidas de este sistema de escritorio son archivos en formato XLS (Microsoft Excel) con:

- La traza, que es el camino que sigue cada registro a través de los vectores teóricos que evalúan la coherencia de la información que contiene, es decir, a qué criterio ingresó, imágenes resultantes y tratamientos aplicados, junto con los cambios realizados a sus variables.
- Los reportes que además de mostrar la información del cuestionario modificado o no por los tratamientos, sirven para la revisión del comportamiento de los criterios.

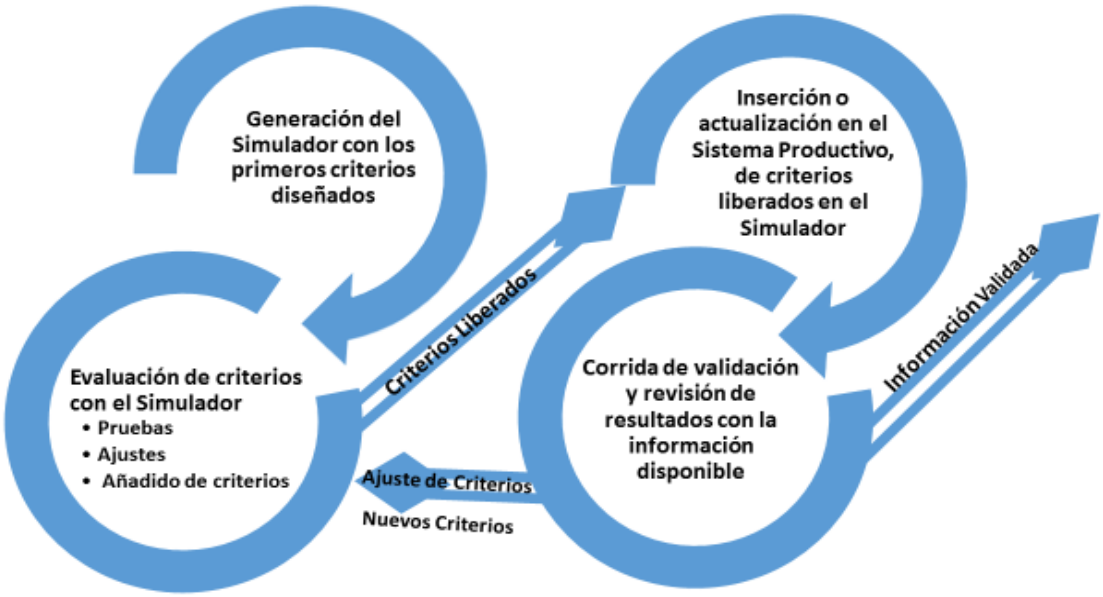


Cabe destacar que una vez que el usuario, a través del uso del simulador, determinaba que un criterio satisfacía enteramente sus requerimientos, las clases Java del criterio y mapeadores asociados, fueron insertadas, **sin cambio alguno de código,** en el sistema de validación en producción.

En lo que respecta al modelo de datos, adicionalmente a los objetos de intercambio de información, fue creado un modelo relacional con el objetivo de que el usuario pudiera consultar, en todo momento, la información validada de viviendas y personas, así como la traza asociada a su proceso de validación. Los registros en este modelo fueron insertados o actualizados desde los mismos objetos de intercambio, cuyos métodos miembro realizaron dichas operaciones. Se crearon también tablas para el uso del sistema, en donde se almacenaron parámetros útiles para la ejecución de la validación -por ejemplo, el número de hilos, grupos y tamaño de la colección que en cada ciclo se procesó-, que para el caso que nos ocupa, fueron 5 grupos de 5 hilos y 50 elementos por colección, de tal manera que corrieron en paralelo 1,250 cuestionarios de viviendas con sus respectivas personas de manera constante.



Finalmente, el flujo de procesamiento utilizado se ve representado como sigue:



Es importante resaltar que mediante este flujo fue posible desarrollar el software asociado a los criterios, a la par de que éstos eran delineados por el usuario, ahorrando tiempo en el proyecto. Otro aspecto significativo a resaltar es que al validar información real en el sistema productivo se encontraron casos atípicos que provocaron la creación de nuevos criterios o ajustes en los existentes. Dichos ajustes y criterios nuevos se programaron y fueron probados en el simulador antes de ponerse en producción, lo que implicó un control de calidad altamente aceptable.

CIFRAS[7] RELEVANTES DE VALIDACIÓN DE LA EIC2015

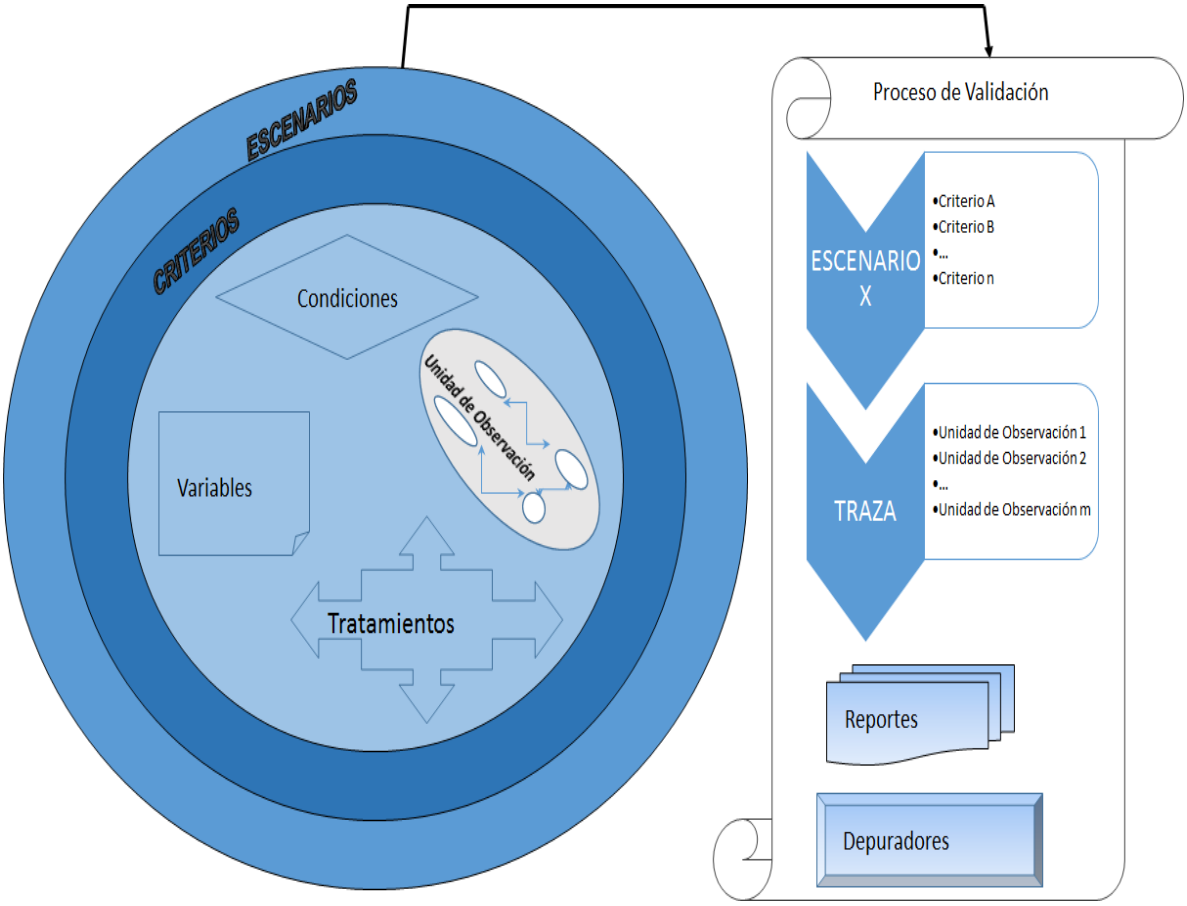
PROCESO DE VALIDACIÓN					
VALIDADAS		TOTAL DE CRITERIOS	INICIO	FIN	OBSERVACIONES
VIVIENDAS	PERSONAS				
5,854,392	22,664,661	240	21/09/2015 16:48	22/09/2015 01:51	Tiempo total 09:03:07 Hrs.
REPORTES DE VALIDACIÓN					
DE ENTRADA-SALIDA	DE CRUCE DE VARIABLES	CORTES DE EDAD	OTROS	TOTAL	OBSERVACIONES
334	161	62	5	562	Todos a nivel municipal, estatal y nacional, formato Excel. Tiempo total de generación 38:45:50 Hrs.

[7] Los valores de viviendas y personas presentados en el cuadro no son cifras definitivas, considerando que como resultado de la validación es posible la baja lógica de registros. Adicionalmente, es una práctica común la realización de operativos de campo para recuperar cuestionarios faltantes.

Propuesta de mejora

La primera propuesta de mejora gira en torno al cambio de paradigma en la forma en la que se estructura el proceso de validación:

Visión Actual



Actualmente el conjunto de unidades de observación, variables, condiciones y tratamientos, conforman los criterios que, en conjunto, forman escenarios para el proceso de validación. Una vez elegido un escenario compuesto por *n* criterios, al ejecutar el sistema se evalúa la información aplicando, en un orden específico, criterio a criterio a cada registro, generándose las trazas correspondientes a *m* número de unidades de observación. A esta ejecución se le asocian los reportes correspondientes, así como a los depuradores que permitirán a los analistas interactuar con la base de datos para corregir las inconsistencias que se presenten.

La propuesta se centra en que a cada unidad de observación, se le asocie un escenario, de tal manera que cada una de ellas tenga como parte de su

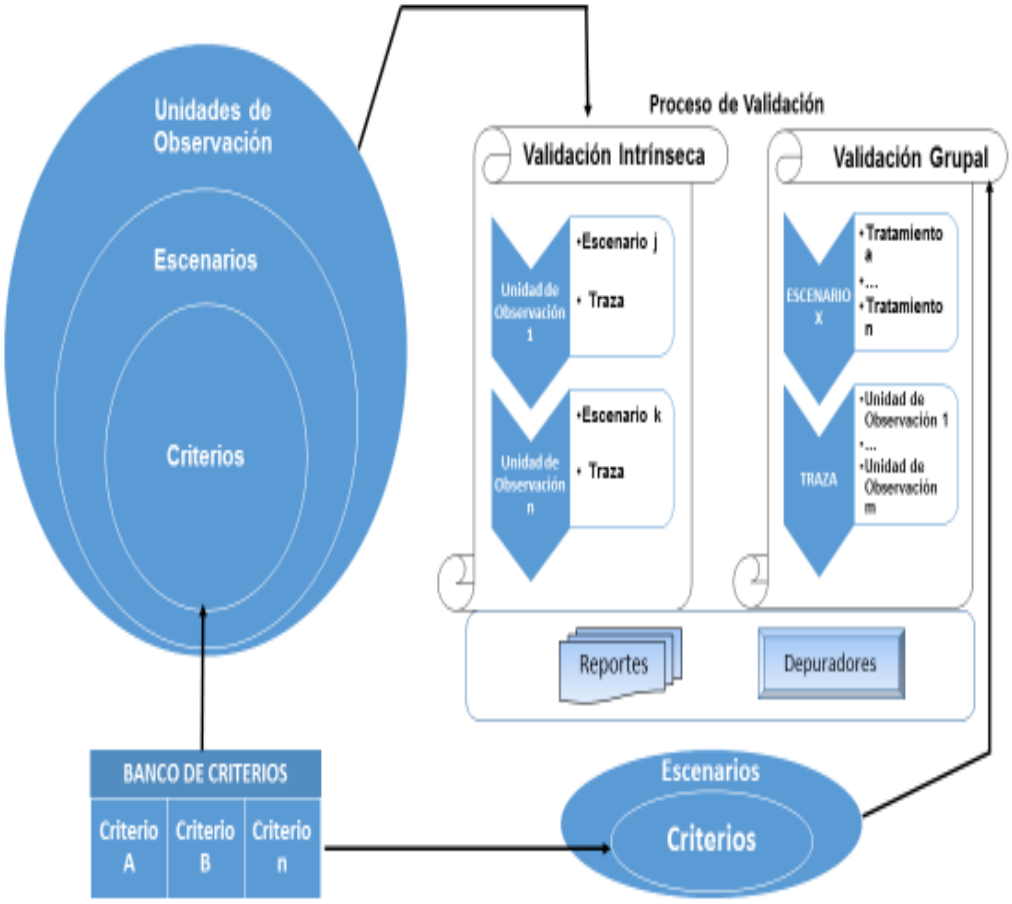
comportamiento la posibilidad de revisar la consistencia y congruencia de su información asociada. Con ello se brindaría la posibilidad de que la validación se realizara en función de conjuntos derivados de clasificaciones lógicas de los datos; esto es, se abriría la oportunidad de aplicar un escenario, es decir, un juego de criterios específicos a un cierto conjunto de información, de la misma manera se aplicarían escenarios distintos a otros conjuntos de unidades de observación. Es importante señalar que este esquema propuesto, también permite el proceso de validación tal y como ahora se realiza, simplemente asociando el mismo escenario a todas las unidades de observación.

Una situación derivada de este paradigma implica que desde el momento del proceso de la carga de datos al esquema de validación, éstos, al ser insertados en su destino, ya contengan la evaluación realizada por los criterios de validación al ser dicha valoración intrínseca a su comportamiento. Para contextualizar esta mejoría se debe señalar que actualmente el proceso de carga (transferencia de información de un esquema de base de datos de captura a uno para el tratamiento de la información), consume un porcentaje considerable del tiempo total del proceso de validación.

La creación de un banco de criterios re utilizables también forma parte de esta oferta; se propone crear, semejante a un juego de herramientas, criterios genéricos que, gracias a la arquitectura creada basada en mapeadores, sea posible su utilización en diferentes escenarios e incluso distintos censos, con la ventaja de que serán artefactos que hayan probado ser correctos y eficaces. Se trata entonces, de la posibilidad de utilizar los mismos criterios en más de un escenario a la vez y, por ende, de que las piezas de software ya construidas sean utilizadas a futuro en los sistemas de validación de censos venideros.

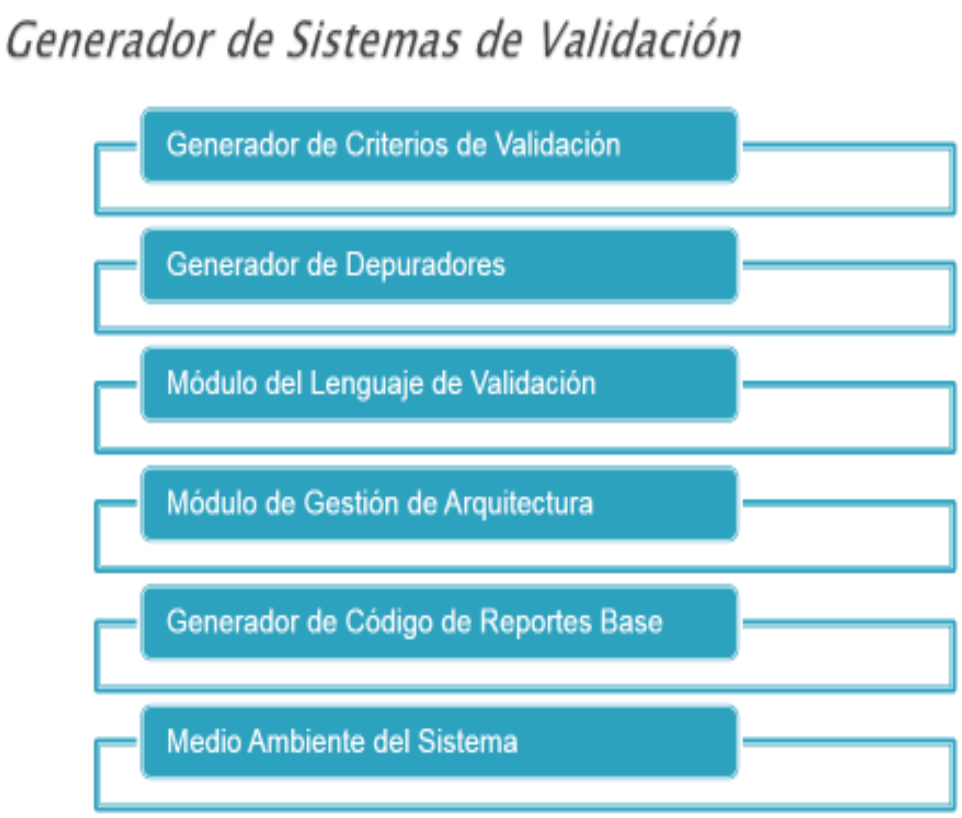
La propuesta también sugiere, como parte de su estructura, otro tipo de validaciones que involucran, a diferencia de la exploración unidad por unidad de observación, la revisión de una vecindad de ellas para aplicársele algún tipo de tratamiento estadístico, a fin de encontrar inconsistencias para ese conjunto de unidades con alguna característica común de interés. Incluso, se tiene la idea de utilizar algún clasificador, que en un futuro cercano, pudiera ser implementando mediante alguna técnica de inteligencia artificial.

A continuación se presenta la estructura propuesta:



Como se planteó en el documento, hemos desarrollado generadores de código para la implementación de criterios de validación con vectores teóricos. Con ello hemos aumentado la velocidad del desarrollo de los sistemas correspondientes, logrando disminuir ostensiblemente sus tiempos de entrega y mejorando la calidad de los mismos.

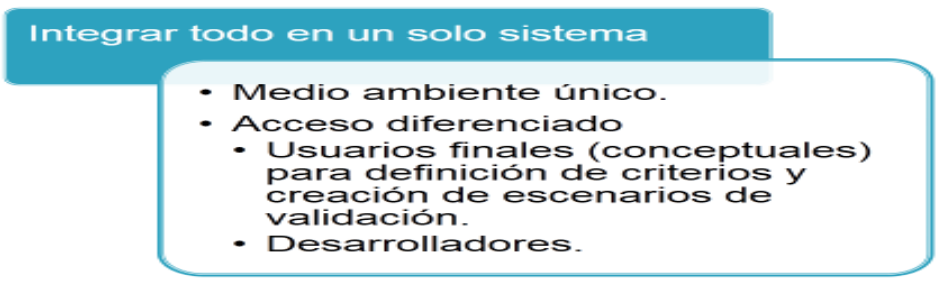
En este sentido, la propuesta de mejora consiste en el desarrollo de un sistema integral de generación de sistemas de validación, respetando la filosofía de proponer la solución tecnológica pensando siempre en que ésta debe responder a las necesidades y requerimientos de las áreas conceptuales, para lo cual se construyen sistemas a la medida basados en el re uso de componentes y trabajando juntos, como un solo grupo, con el área requiriente.



Los pasos a seguir son:

- Desarrollar interfaces para los generadores de código ya construidos.
- Desarrollar el Generador de Depuradores.
    - *Masivos*. A través de archivos de actualizaciones.
    - *Unitarios*. Que pueden ser referidos a un criterio en particular (únicamente las variables que intervienen), o bien, depuradores generales (todas las variables); considerando siempre presente la vista completa del cuestionario para consulta de los datos a depurar.
    - Construcción del control de cambios respectivo.
    - Desarrollo de interfaces del generador.
- Módulo del Lenguaje de Validación.
    - Crear el lenguaje de validación.
        - Será un lenguaje formal.
        - Servirá para describir las condiciones y tratamientos dentro de la tecnología de vectores teóricos de forma clara, precisa, estructurada, ordenada y conocida por la comunidad implicada.

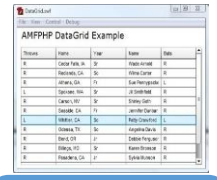
  - Se construirá un compilador que tome como entrada un texto escrito en el lenguaje de validación y produzca como salida clases Java listas para su ejecución.
  - Se utilizarán las interfaces desarrolladas para definir criterios y, a través del compilador, se producirá código útil ejecutable, es decir, se generarán los módulos de evaluación de criterios desde la captura misma del requerimiento.
- Desarrollar el Módulo de Gestión de Arquitectura y Control.
  - Generador del modelo de datos.
    - Interfaces para su definición.
  - Desarrollo de interfaces para la administración de direcciones y conexiones a servidores productivos, de aplicaciones y de base de datos.
  - Construcción del controlador de procesos.
- Desarrollar el Generador de Código de Reportes Base.
  - De matrices de entrada-salida.
  - De avance.
  - Listado de cuestionarios inconsistentes.
  - Por variable: valores, omisiones y porcentaje de cambios.
  - De cruce de variables.
- Integración del sistema
  - Conjunción de los módulos que permitirán:
    - A partir de la definición de variables de un cuestionario, diseñar el modelo de datos asociado y sus objetos de intercambio de información Java-Oracle.
    - La creación, conformación y control de ejecución (*multihilos*), de todos los criterios que formen la validación.
    - La generación del simulador para la prueba de los criterios diseñados.
    - La generación de las estructuras y programas para la producción de reportes básicos.
    - La creación de la estructura de código del proyecto, que será la que se publique en los servidores.



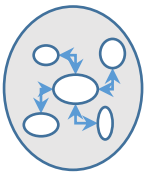
Flujo del Sistema Integral de Generación de Sistemas de Validación

1 Carga de Variables del Cuestionario

2 Alta de Proyecto



3 Definición de Atributos del Proyecto

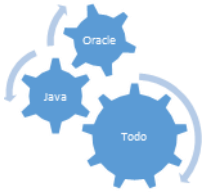


4 Modelado de la Unidad de Observación

$$G_5(X_1, X_2, X_3, X_4) = 8X_1 + 4X_2 + 2X_3 + X_4 + 1$$

IMAGEN	TRATAMIENTO
1	Correcto, C_05 =1
6	
11	
16	
2	Si $B_i = 0$, hacer $K_{630}\_A = K_{630}\_A + K_{610}\_A$ Correcto, C_05 =1 En otro caso DM, C_05 = 3

5 Definición de Criterios



6 Generación de Código

Conclusiones

Aunque todos los censos comparten características comunes, cada uno tiene sus particularidades, incluso siendo del mismo tema -los Censos Económicos 2009 fueron distintos de los Censos Económicos 2014- y, a pesar de que al final se desea medir básicamente lo mismo, difieren desde la forma del operativo de campo para el levantamiento de la información, hasta los aspectos conceptuales de los criterios de validación y explotación. Esto se explica por el dinamismo del país: cambia la idiosincrasia, la forma de hacer las cosas, los usos y costumbres, así como la tecnología con sus vertiginosos avances. Por ello en cada nuevo censo es necesario adecuar la manera en que se abordan la recopilación, procesamiento, explotación de los datos así como entrega de resultados.

Lo descrito en el párrafo anterior me permite concluir que la mejor opción no es desarrollar un sistema general de validación, sino uno que genere el código específico para cada censo, siempre tomando ventaja de las tecnologías emergentes que mejor se adapten a la solución. En este sentido la industria del software repetidamente señala el uso de las llamadas mejores prácticas. No obstante, en mi concepto, mejores prácticas es un tema creado inicialmente con objetivos de mercadeo y asociado a la solución mágica a todos los problemas; supone que, de todas las opciones de solución, es la mejor. Sin embargo, mi experiencia me indica que la mejor solución no necesariamente está guiada por las mejores prácticas comunes y, en cambio, siempre coincide con aquella que considera ampliamente el contexto. Adicionalmente, desarrollar software no es un asunto meramente manufacturero, en realidad, es una combinación de ciencia, ingeniería, arte y administración de proyectos. Por ello la simplificación en la construcción de los sistemas de validación, desde mi punto de vista, reside en combinar los elementos asociados específicamente al dominio del evento estadístico en cuestión, encapsulando el conocimiento del negocio así como técnico asociado, y contando con programadores de gran habilidad para relacionarse con otras disciplinas, mismos que ofrezcan soluciones adecuadas donde todo lo repetible lo automaticen.

Muy importante es reunir a un conjunto de desarrolladores de alta calidad y comprometidos con los proyectos, tarea nada sencilla y que en el caso del grupo que me honro en coordinar, ha sido una labor de al menos ocho años en el que, como responsable de una área de desarrollo de sistemas, he trabajado duro formando al personal, tanto en la parte técnica como en la conceptual, destacando el favorecimiento de la acumulación de experiencia única e invaluable, toda vez que el quehacer de desarrollar sistemas para los censos es muy especial. Esto lo digo desde mi óptica de muchos años dedicados a la construcción de todo tipo de sistemas informáticos, desde aquellos de escritorio para agentes aduanales, hasta enormes desarrollos empresariales para PEMEX y la SEDENA. Los sistemas de validación censal, a diferencia de otros de mayor

continuidad con procesos periódicos de mejora, nacen y mueren en cada censo, teniendo un período de vida corto desde su desarrollo hasta su producción, debiendo ser altamente precisos y veloces.

Cabe señalar que una consideración importante es que, si bien se busca que cada vez el usuario dependa menos de los desarrolladores, durante el tiempo destinado al proceso de validación siempre habrá, y así debiera ser, una muy fuerte interacción de los programadores con el área que delinea el requerimiento. A pesar de contar con generadores de código maduros, seguirá siendo necesario destinar recursos humanos para el diseño y programación de módulos adicionales, así como ajustes a los ya existentes. De la misma forma se requerirán programadores para atender la solución de situaciones no previstas.

En lo que respecta al intercambio de información entre la aplicación y la base de datos, está demostrado que fue un gran acierto generar nuestra propia capa de persistencia, pues nos permitió lograr las velocidades esperadas de lectura y escritura. El poder leer o escribir colecciones completas de unidades de observación en una sola llamada, no sólo reduce ventajosamente a una, en vez de muchas, el número de conexiones a la base de datos, sino que también nos permite, en un solo paso, un manejo más eficiente y controlado del trabajo en múltiples hilos de procesamiento.

En referencia a la formación profesional que me brindó mi querida Facultad de Ciencias, debo argumentar que sin ésta me hubiera sido imposible haber diseñado e implementado la solución presentada. Y que gracias a dicha formación y a la oportunidad que me brinda el INEGI, trabajo en lo que tanto me apasiona: el desarrollo de software. Sin falsa modestia, me permito comentar que personalmente programé, entre otros componentes, los generadores de código de los objetos en la base de datos, y que la arquitectura de la aplicación, así como la visión futura del sistema en su conjunto, fue ideada y diseñada por quien esto escribe.

Finalmente, es la formación profesional de actuario la que me ha permitido poner mi granito de arena en la consecución de algunos logros obtenidos por el INEGI en los diferentes censos sucedidos desde 2004. Repito las palabras que dije a mis colaboradores al finalizar la ronda censal en 2010: “…cuando escuchen en la radio, o vean en la televisión o en Internet que anuncian los resultados de un evento tan grande como es un censo, deben estar orgullosos, porque también fue gracias a ustedes que se lograron”.

México DF, Febrero de 2016.