Distr.
RESTRINGIDA

E/CEPAL/R. 338
27 de mayo de 1983

ORIGINAL: ESPAÑOL

C E P A L

Comisión Económica para América Latina

PRINCIPIOS DE UN SISTEMA INTEGRADO DE PROCESAMIENTO MUESTRAL

Este documento fue preparado como versión preliminar por el señor Carlos Cavallini, Asesor Regional en Muestreo para Encuestas y Censos de Población. Las opiniones expresadas en este trabajo son de exclusiva responsabilidad del autor y pueden no coincidir con las de la Organización.

83-6-828

PRINCIPIOS DE UN SISTEMA INTEGRADO DE PROCESAMIENTO MUESTRAL

(versión preliminar)

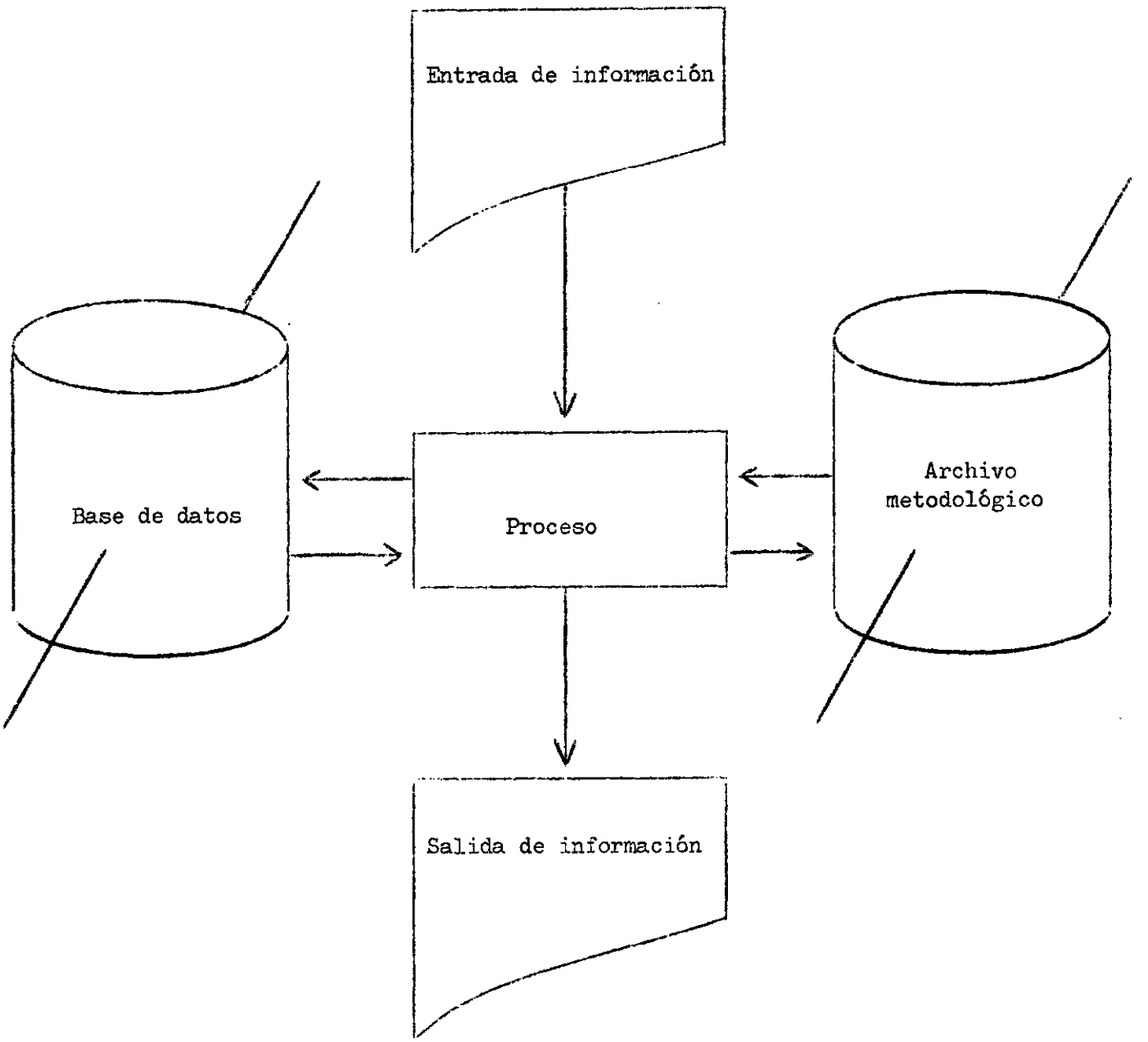
1. El Sistema Integrado de Procesamiento Muestral, SINPROM, pretende ser un conjunto de reglas y principios sobre muestreo estadístico, entrelazados entre sí, que permitan, en forma sencilla y fácilmente entendible, mantener coordinada la información y la metodología que se aplica en las distintas investigaciones muestrales que realizan las Oficinas Nacionales de Estadística. De ahí el nombre de Sistema Integrado. El procesamiento del Sistema puede ser manual, mecánico o ambos a la vez. Pero dado que todas las Oficinas ya cuentan, o están en vías de adquirir un sistema moderno de computación, el enfoque se hará teniendo en cuenta esta alternativa.

La utilidad del Sistema tendrá significación, sobre todo, en los programas de encuestas, dado que permitirá mantener la coordinación de los programas, con el consiguiente beneficio de obtener una integración mayor de información y por ende, de resultados. Esta integración se podrá, además, llevar a cabo a través del tiempo y del espacio, por ejemplo, con la construcción de series cronológicas, ya sea por estratos geográficos o por dominios de estudio.

2. Es de hacer notar que actualmente los tiempos de procesamiento interno de una computadora digital se miden en manosegundos. Los tiempos de impresión en papel ya llegan, en laboratorio, a 70 000 líneas por minuto. La densidad de archivo en un disco de tamaño normal es de alrededor de 500 millones de caracteres. El problema persiste en la entrada de información, que comparado con los otros tiempos, es significativamente lenta. La parte relevante para el Sistema reside en la densidad de los archivos. Por ejemplo, en un país de 10 millones de habitantes, si un censo de población recoge 50 caracteres por habitante, es posible almacenar dicho censo en un disco. Se podrá así llegar a mantener en línea más de un archivo por unidad de máquina, por ejemplo, por unidad de discos, lo cual será de utilidad para aquellas Oficinas que cuenten con un sistema pequeño de computación.

/3. La

3. La comunicación del Sistema, en su forma amplia adopta la siguiente lógica:



/4. La Base

4. La Base de Datos contendrá los datos de la información. El Archivo Metodológico contendrá los modelos estadísticos y matemáticos, y los programas de utilización.

5. El diseño del Sistema estará regido por los siguientes criterios generales:

i) En la información de las bases y de los archivos primará un criterio eutáxono, ordenamiento y correlación, y de almacenamiento acumulativo: el sistema se incrementa a medida que se desarrolla.

ii) En los archivos se utilizarán los algoritmos de funciones generatrices.

iii) La inter-relación base-archivo condicionará la evaluación de las funciones.

iv) En la programación para el acceso, ejecución y salida del sistema deberá prevalecer el criterio de simplicidad.

v) La base y el archivo tendrán acceso directo y serán independientes de la programación, y viceversa, "ramdom access".

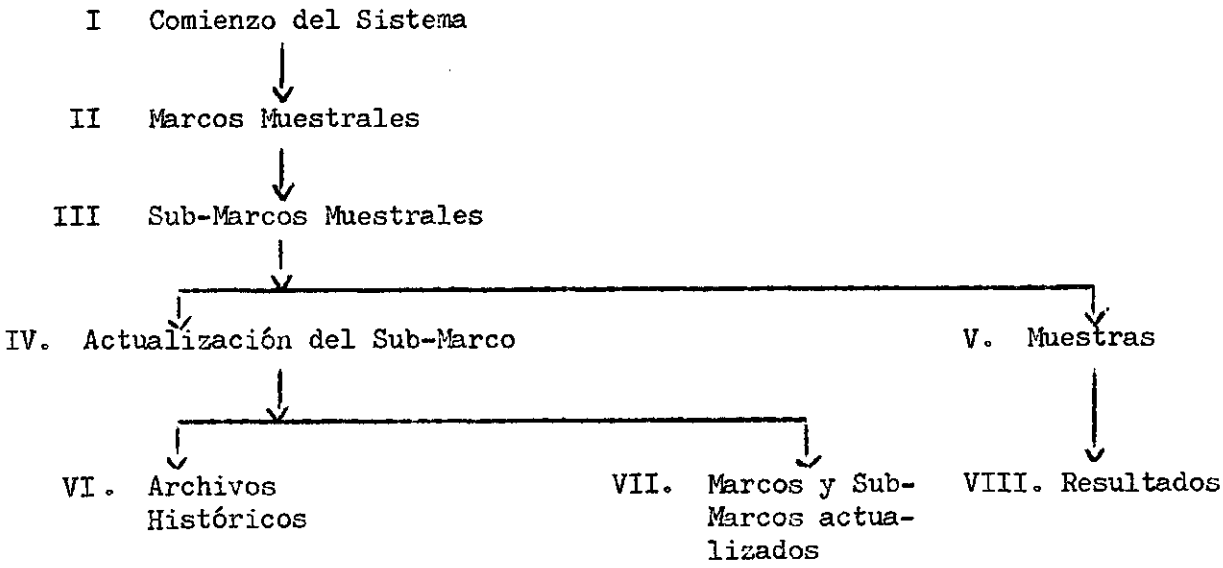
vi) Cada base, archivo, programa, función, modelo, etc. tendrá un nombre particular, biunívoco, que se registra en un diccionario con la descripción, concepto, definición, tamaño de registro, etc. de dicho elemento.

Generación del Sistema

6. En función de la infraestructura existente y de la configuración que se adopte en dimensionar al Sistema, se establecerán las reglas que regirán al mismo. El punto de inicio será, por tanto, la construcción de un "Sistema vacío".

7. Teniendo en cuenta la situación promedio de las Oficinas Nacionales de Estadística de la región, para el llenado del sistema se puede sugerir el siguiente orden de comienzo:

/I Comienzo del



I. Comienzo del sistema. Se definirá al sistema vacío.

i) Equipo de procesamiento
   a) Dimensión del equipo
   b) Unidades para la formación de Archivo
   c) Unidades para la formación de Bases

ii) Programación
   a) Lenguajes
   b) Programas de biblioteca
   c) Programas genéricos

iii) Cartografía
   a) Organización

iv) Establecer las normas básicas que regirán al sistema
   a) Accesibilidad (random access)
   b) Flexibilidad
   c) Asociación (distintos trabajos)

v) Diccionario

/El diccionario

El diccionario abarcará todos los elementos del sistema con una descripción detallada de los mismos, conjuntamente con los diagramas en bloque, diagramas en detalle, programas de instrucciones, programas en simbólico, programas en absoluto, etc.

II. Marcos Muestrales. El sistema trabajará bajo el supuesto de marcos finitos.

i) Definición de las unidades muestrales

a) Areas geográficas políticas

. Departamentos, Provincias, Municipios, Distritos, Ciudades, Pueblos, etc.

b) Areas geográficas no políticas

. Unidades geo-estadísticas, sub-unidades geo-estadísticas, segmentos rurales, segmentos urbanos, manzanas, etc.

c) Características de las unidades

. Viviendas, hogares, personas, factores de accesibilidad, indicadores estadísticos, indicadores socioeconómicos, etc.

d) Código de unidad

. Ubicación geográfica

. Area

. Regiones

. Dominio de Estudio

ii) Entre los factores de accesibilidad se pueden mencionar, sobre todo para el área rural, distancia en kilómetros desde el centro poblado más cercano para llegar a la unidad, tiempo en horas para recorrer esa distancia, estado de los caminos, medios de transporte, épocas de transitabilidad, costo del transporte, tipos de hospedaje, costo del hospedaje, cota de la unidad, temperatura, etc.

iii) Con respecto a los indicadores estadísticos, algunos se podrán ir construyendo a medida que se vayan llevando a cabo las investigaciones. Entre ellos se pueden citar, por ejemplo, la correlación intraclase para determinadas variables, coeficientes de relación entre distintos atributos, etc.

/iv) Entre

iv) Entre los indicadores socioeconómicos, se pueden citar, porcentaje de obreros, tasas de hacinamiento, coeficientes de instrucción, tipo de producción económica predominante, etc.

Estos valores permitirán la construcción de estratos, con el consiguiente aumento de eficiencia por unidad de costo.

v) El código de la unidad reviste singular importancia, dado que a cada unidad deberá corresponderle un solo código y cada código identificará una sola unidad, es decir, habrá correspondencia biunívoca entre los elementos de cada conjunto. El código de la unidad será un componente de subcódigos, con la ubicación geográfica, según la división que tenga el país, el área, la región, el código según los estratos y subestratos que se adopten, el código de los dominios de estudio, etc. Este código de la unidad se podrá ir desarrollando en función de las necesidades que se vayan presentando. La diferencia entre el código de la unidad y las características de las unidades, es que si bien ambos sirven para estratificar, aquél es fijo mientras que éstas pueden variar.

vi) De acuerdo con el planteo básico del Sistema, el primer paso conduce a la construcción de los marcos muestrales.

En un caso particular, un marco muestral será el censo de población. En la mayoría de los casos el mismo estará compuesto por unidades muestrales que corresponderán a áreas geo-estadísticas. Las condiciones principales de estas unidades deben ser a) no estar traslapadas; b) límites identificables no imaginarios, no variables y de fácil reconocimiento en el terreno; c) tener una relación área-población que sea práctica para trabajar; d) factible de ser divididas en sub-áreas geo-estadísticas; e) adicionando estas áreas se obtendrá la población de estudio.

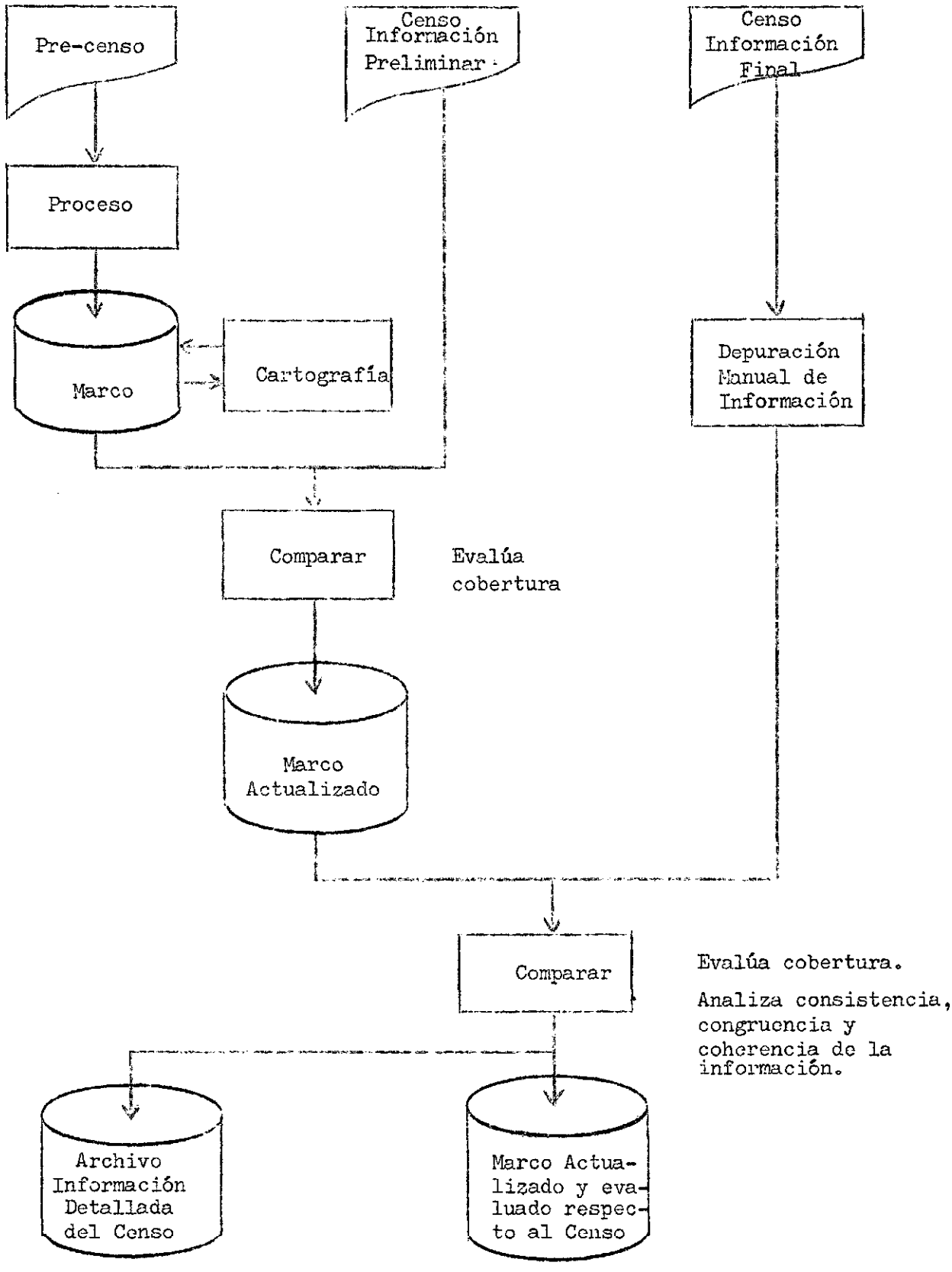
Generalmente el pre-censo con la correspondiente cartografía provee este marco que servirá para el censo e investigaciones por muestreo.

El censo proveerá las características y los factores de costo para ponderar y para estratificar el marco.

Un diagrama en bloque para construir el marco podrá adoptar la siguiente lógica.

/Diagrama en

Diagrama en Bloque para la Construcción de un Marco



III. Sub-marcos muestrales.

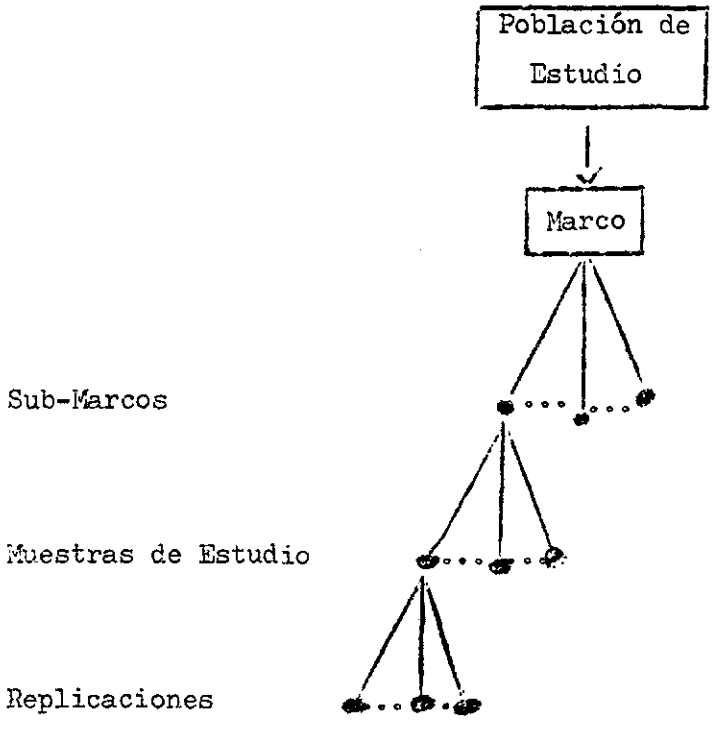
i) Construído los marcos que representarán a las poblaciones de estudio, el sistema deberá prever la confiabilidad y la actualización periódica de los mismos.

La alternativa práctica es la formación de sub-marcos, división del marco en grupos aleatorios, cada uno con similar grado de representatividad.

Un análisis de las variaciones entre los sub-marcos con respecto a las variaciones dentro de los sub-marcos, permitirá conocer la confiabilidad de estos sub-marcos y adoptar una decisión en la selección de un sub-marco para mantener actualizado y servir de generador de muestras.

Este sub-marco seleccionado, muestra del marco, será a su vez dividido en sub-muestras. Una de estas sub-muestras, o varias sub-muestras adicionadas, formarán las muestras de estudio. Para aumentar la eficiencia tanto analítica como práctica, estas muestras así construídas podrán ser replicadas.

El modelo jerárquico de desarrollo es el siguiente,



ii) Para la formación de los sub-marcos o muestras maestras, se pueden adoptar distintos criterios. Uno puede ser, por ejemplo, dividir al marco por estrato y dentro de cada estrato, en función de algún criterio de tamaño de muestra maestra, construir los sub-marcos. Otro puede ser la división en función de algún tamaño máximo de muestra, como pueden darlo las investigaciones demográficas, etc.

iii) Para conocer la confiabilidad del censo, y por tanto del marco construído, con respecto a la población de estudio, se utilizarán alguno de los métodos tradicionales de evaluaciones censales, por ejemplo, métodos analíticos, métodos directos, de registro, de autoevaluación, etc.

Es de hacer notar que la evaluación de la cobertura de áreas del censo ya lo provee el Sistema, mientras que la evaluación de la cobertura de personas y de la confiabilidad de la información se podrá realizar utilizando alguna de las muestras de estudio.

iv) Cada una de las divisiones seleccionadas, muestras de estudio y replicaciones, se almacenarán en la Base de Datos de acuerdo con las normas que se establezcan, las cuales se regirán por, a) acceso inmediato a la información; b) formación de grupos basados en los métodos de aleatorización, estratificación y replicación; c) asociación de la información dentro de las bases, dentro de los archivos y, entre bases y archivos; d) la construcción y el desarrollo de un diccionario.

v) A medida que se vayan desarrollando nuevos trabajos, nuevas técnicas, nuevos modelos, nuevas experiencias, el sistema se irá autoabasteciendo de los mismos así como de los resultados que se vayan logrando.

vi) Dado que el sistema poseerá la mayoría de las funciones estadísticas de aplicación corriente con los correspondientes valores de distribución, el mismo podrá ser consultado en forma independiente en la aplicación de "tests" estadísticos, de pruebas de significación, de análisis de las variaciones, de análisis factoriales, etc.

/IV. Actualización

IV. Actualización del Sub-Marco.

Con respecto a la actualización periódica del sub-marco seleccionado se podrán utilizar, por ejemplo, las distintas investigaciones que se realicen, los conocimientos por denuncia, los estudios que se hagan en cartografía y las investigaciones ad-hoc.

V. Muestras.

Para la conformación de muestras, generalmente se deben tener en cuenta los siguientes puntos,

i) Objetivos de medición
   a) variables
   b) atributos
   c) características

ii) Funciones generatrices, definiciones
   a) tamaño
      . $n = f\ (p;\ t;\ e;\ d;\ v;\ a;\ b;\ c;\ .....)$

n	tamaño muestral
p	objetivo de medición
t	confiabilidad deseada
e	error absoluto aceptado
d	error relativo aceptado
v	coeficiente de variación aceptado
a	condición paramétrica
b	condición paramétrica
c	costo

   b) momentos
      . naturales

$$m_k = E\ x^k$$

      . centrados

$$\mathbf{m}_k = E\ (x - \bar{X})^k$$

      . reducidos

$$q_k = E\left(\frac{x - \bar{X}}{\sigma}\right)^k$$

/k grado

k grado del momento

x variable

$\bar{X}$ media parámetro

σ desvío estándar

E esperanza matemática

c) intervalos de selección, i,

$$i = f(N;n)$$

N tamaño del marco

d) arranque aleatorio, g,

$$1 \leq g \leq i$$

iii) Funciones generatrices, programas

a) n

b) m_k ; n_k ; q_k

c) i

d) g (tabla aleatoria o generación de números aleatorios)

iv) Uno de los primeros trabajos a realizar será la selección de muestras. Para ello el sistema contará con un programa que, dada la variable de estudio, el error y la confiabilidad deseada, generará el tamaño muestral e imprimirá las unidades seleccionadas.

El programa realizará, por ejemplo, los siguientes pasos:

a) Almacena, por programa, el código de la variable de estudio en nuestro caso p, ver V ii)a), los valores t, e y los códigos de las unidades de selección, por ejemplo, las unidades de selección primaria podrán ser las áreas geo-estadísticas y las unidades de selección secundaria podrán ser las viviendas;

b) Evalúa p, a, b y n, y los imprime

c) Evalúa i

d) Genera g, (g + i), (g + 2 i), ..., $\left[g + (m - 1)\, i \right]$

e) Imprime los códigos de las unidades de selección primaria que han sido seleccionadas.

VI. Archivos

VI. Archivos históricos. Los archivos históricos estarán integrados a la asociación del sistema.

i) Programas de generación de los Archivos históricos y de consulta.

VII. Marcos y sub-marcos Actualizados.

i) Programas de actualización.

VIII. Resultados.

i) Programas

a) De impresión de listas

b) Tabulados y cuadros

. Muestrales y expandidos

8. Modelos estadísticos.

El primer modelo lineal estadístico aditivo generatriz a archivarse adoptará la forma

$$y = 0 + \overline{\sum} \; t + e$$

donde

y observación

0 parámetro

t tratamiento o factor

e error aleatorio