

PACKAGE “SAMPLINGSTRATA”

UN RECURSO PARA:

- ESTRATIFICAR UN MARCO MUESTRAL
- SELECCIONAR UNA MUESTRA CON PROPÓSITOS MÚLTIPLES,
- CON RESTRICCIONES DE PRECISIÓN Y COSTOS FIJOS

TAMAÑO Y ASIGNACION EN EL CASO MULTIVARIADO

- **Máxima:** “Una muestra es aceptable si los errores muestrales están por debajo de los límites predefinidos y los costos son sustentables”
- **Propuesta:** Sea “U” una población con “M” variables auxiliares (X) en condiciones de estratificarla y restricciones de precisión sobre “G” variables objetivos (Y) se busca en forma conjunta:
 - a) La mejor estratificación, o sea, la mejor partición de U
 - b) El menor tamaño de muestra y una asignación que permitan satisfacer las restricciones de precisión.
 - c) Al menor costo posible

ALGORITMO DE BETHEL

UNA ALTERNATIVA PARA BRINDAR SOLUCIONES
CUANDO LA ESTRATIFICACIÓN DE LA POBLACIÓN ES
DADA (H) :

- b) TAMAÑO DE MUESTRA Y ASIGNACIÓN
- c) MINIMIZACIÓN DE COSTOS

Ejemplo típico de un Marco Estratificado con Múltiples Variables Objetivos

	RAZON.SOCIAL	VMP	VENTAS	PO	TIPO	ACTIV	REG
1	VACION SOCIEDAD ANONIMA	1592778	2271	91	3	3	1
2	BUDDENSIEG S A ARGENTINA I C I Y F	1400547	6084	118	2	4	1
3	MAMPER S A	682373	2320	8	3	2	1
4	PACUCA SA	5548441	158	194	2	5	1
5	AGROPECUARIA LA MARIA PILAR SOCIEDAD ANONIMA	2570007	257	200	2	5	1
6	GEN AVE S A	5047769	2053	191	3	5	1
7	GOJAM SCA	4738775	613	224	2	1	1
8	MIROLU S A	3530302	5461	123	3	1	1
9	SANTOS GENCHI S A C I A E I	800604	3134	174	1	2	1
10	NAHUEL MAPA SCA	1307544	7102	96	2	2	1
11	RINCON DE CORRIENTES SA	1347371	2602	210	3	2	1
12	GRANAR S A COMERCIAL Y FINANCIERA	4616832	1853	15	3	4	1
13	VANFLA SOCIEDAD ANONIM	5573398	858	186	2	2	1
14	DOSALO SOCIEDAD ANONIM	1011076	5495	72	1	1	1
15	ESTAR S A AGRICOLO GANADERA COMERCIAL E INMO...	1886855	3417	118	3	3	1
16	CAPDEVILA JUANA MARIA MORALEJO	26160	7949	85	3	2	1
17	CHANCHI HUE S A	1197947	2258	206	1	4	1

dataframe “Marco.Empresas.RData”

#U=4782 unidades

Planteo,

Variables Objetivos :

$$Y_1 = PO \quad Y_2 = VENTAS \quad Y_3 = VMP$$

Variables Estratificadoras :

$$X_1 = REG(3) \quad X_2 = ACTIV(5)$$

Estratificación (15 estratos) dado por el cruce de REG (3) y ACTIV(5)

Se desea estimar PO, VENTAS y VMP (o sea G=3) con precisión del orden de 3%, 5% y 5% en términos de CV respectivamente a nivel de U

$$\min f(n_1, \dots, n_H) = \min \left\{ C_0 + \sum_{h=1}^H c_h n_h \right\}$$

sujeto a $\begin{cases} CV_{PO} \leq 0.03 \\ CV_{VENTAS} \leq 0.05 \\ CV_{VMP} \leq 0.05 \end{cases}$

Recordar Planteo con una Restricción . . .

Dados un costo inicial C_0 , costos fijos por estrato c_1, c_2, \dots, c_H , un valor V_0 (cv_0) para la varianza (coef. de variación) de $\hat{T}_{\pi y}$ la asig. óptima viene de:

$$\min \left\{ C_0 + \sum_{h=1}^H c_h n_h \right\} \text{ sujeto a: } \sum_{h=1}^H W_h^2 \left(\frac{1}{n_h} - \frac{1}{N_h} \right) S_{yh}^2 = V_0 \quad (= cv_o^2 T_Y^2)$$

$$n_h^{opt} = \frac{\left(W_h S_{yh} / \sqrt{c_h} \right) \left(\sum_{h=1}^H W_h S_{yh} \sqrt{c_h} \right)}{cv_0^2 T_Y^2 + \sum_{h=1}^H W_h^2 S_{yh}^2 N_h^{-1}}$$

$$n = \sum_{h=1}^H n_h^{opt}$$

Algoritmo de Bethel para más de una Restricción (Resolución Iterativa)

Dada una estratificación sobre U, una función lineal de costos y un conjunto de restricciones en términos de CV sobre las G variables objetivo, se busca minimizar

$$f(n_1, \dots, n_H) = C_0 + \sum_{h=1}^H c_h n_h \quad \text{sujeto a} \quad CV_{yg}^2 \leq cv_g^2 \quad g = 1, \dots, G$$
$$\text{con} \quad CV_{yg}^2 = \frac{\sum_{h=1}^H N_h^2 \left(\frac{1}{n_h} - \frac{1}{N_h} \right) S_{ygh}^2}{T_{yg}^2} \Rightarrow \frac{\sum_{h=1}^H N_h^2 \left(\frac{1}{n_h} - \frac{1}{N_h} \right) S_{ygh}^2}{T_{yg}^2 cv_g^2} \leq 1$$

y cv_g dados para cada variable objetivo

Linealizando las Restricciones . . .

Transformando el problema para minimizar una función convexa con restricciones lineales

Haciendo $z_h = \begin{cases} 1/n_h & \text{si } n_h \geq 1 \\ \infty & \text{en otra caso} \end{cases}$

Objetivo : $\begin{cases} \min f(z_1, \dots, z_H) = C_0 + \sum_{h=1}^H \frac{c_h}{z_h} \\ 0 < \sum_{h=1}^H a_{hg} z_h \leq 1 \quad g=1, \dots, G \end{cases}$ con $a_{hg} = \frac{N_h^2 S_{y_{hg}}^2}{T_{y_g}^2 c v_{y_g}^2 + \sum_{h=1}^H N_h S_{y_{gh}}^2}$

- ⊗ Resolución numérica por cualquier método NLP
- ⊗ Puede no haber solución
- ⊗ Si el algoritmo converge, lo hace a la solución

Solución por SamplingStrata, bethel()

Dos comandos y 3 dataframes

buildStrataDF(dataframe)

bethel(objname1,objname2,printa = TRUE)

objname1 y *objname2* son del tipo *dataframe*

⊗ Por simplicidad se asume $C_0 = 0$ y $c_h = 1$, $\forall h, h = 1, 2, \dots, H$

```
> library("SamplingStrata")
```

Paso 1:

Crear las Variables:

Y's (objetivos)

X's (estratificadoras)

y

domainvalue (dominios)

```
Marco=Marco.Empresas
```

```
#####
# Variables Y sobre la que se busca
# precision en la optimizacion
#####
```

```
Marco$Y1=Marco$PO
Marco$Y2=Marco$VENTAS
Marco$Y3=Marco$VMP
```

```
#####
# Variables X que definen la
# estratificacion en la optimizacion
#####
```

```
Marco$X1=Marco$REG
Marco$X2=Marco$ACTIV
```

```
# Variable de control necesaria
# para la optimizacion
Marco$domainvalue=1
```

dataframe Marco con las nuevas variables

RAZON.SOCIAL	VMP	VENTAS	PO	TIPO	ACTIV	REG	Y1	Y2	Y3	X1	X2	domainvalue
VACACION SOCIEDAD ANONIMA	1592778	2271	91	3	3	1	91	2271	1592778	1	3	1
BUDDENSIEG S A ARGENTINA I C I Y F	1400547	6084	118	2	4	1	118	6084	1400547	1	4	1
MAMPER S A	682373	2320	8	3	2	1	8	2320	682373	1	2	1
PACUCA SA	5548441	158	194	2	5	1	194	158	5548441	1	5	1
AGROPECUARIA LA MARIA PILAR SOCIEDAD ANONIMA	2570007	257	200	2	5	1	200	257	2570007	1	5	1
GEN AVE S A	5047769	2053	191	3	5	1	191	2053	5047769	1	5	1
GOJAM SCA	4738775	613	224	2	1	1	224	613	4738775	1	1	1
MIROLU S A	3530302	5461	123	3	1	1	123	5461	3530302	1	1	1
SANTOS GENCHI S A C I A E I	800604	3134	174	1	2	1	174	3134	800604	1	2	1
NAHUEL MAPA SCA	1307544	7102	96	2	2	1	96	7102	1307544	1	2	1
RINCON DE CORRIENTES SA	1347371	2602	210	3	2	1	210	2602	1347371	1	2	1
GRANAR S A COMERCIAL Y FINANCIERA	4616832	1853	15	3	4	1	15	1853	4616832	1	4	1
VANFLA SOCIEDAD ANONIM	5573398	858	186	2	2	1	186	858	5573398	1	2	1
DOSALO SOCIEDAD ANONIM	1011076	5495	72	1	1	1	72	5495	1011076	1	1	1
ESTAR S A AGRICOLO CANADERA COMERCIAL FINMO	1886855	2417	118	3	3	1	118	2417	1886855	1	3	1

Estratificación (15 estratos) dado por el cruce de REG (3) y ACTIV(5)
3 Características objetivo, PO, VENTAS y VMP

Paso2:

Creación
de
objname1
con
“buildStrataDF”

$$a_{hg} = \frac{N_h^2 S_{y_{hg}}^2}{t_{y_g}^2 cv_{y_g}^2 + \sum_{h=1}^H N_h S_{y_{gh}}^2} = f(N_h, \bar{Y}_{hg}, S_{y_{hg}})$$

```
#####
# Creacion del dataframe con
# el resumen por estrato de los
# promedios y sd de cada variable Y
#####
```

```
StratResumen=buildStrataDF(Marco)
```

objname1

Resultado del comando “buildStrataDF”

STRATO	N	M1	M2	M3	S1	S2	S3	COST	CENS	DOM1	X1	X2
1*1	257	126.3930	3915.428	1304743.8	72.20618	2388.266	4143005	1	0	1	1	1
1*2	281	124.7794	4097.342	908154.7	73.15509	2384.987	1794076	1	0	1	1	2
1*3	305	122.9967	3984.223	1006462.2	72.79126	2350.350	3171672	1	0	1	1	3
1*4	266	126.5977	3790.173	1029411.3	72.03539	2355.482	3044145	1	0	1	1	4
1*5	272	135.1801	3958.276	982938.1	67.90668	2322.701	1981940	1	0	1	1	5
2*1	360	127.7972	3947.406	679325.8	70.42877	2336.979	1188628	1	0	1	2	1
2*2	324	129.5988	3951.694	768263.6	74.87244	2342.474	1387096	1	0	1	2	2
2*3	304	126.3026	4040.645	617368.3	72.21223	2315.137	1102061	1	0	1	2	3
2*4	336	135.6577	3854.354	706093.5	69.74980	2205.686	1242219	1	0	1	2	4
2*5	308	126.6916	4090.662	779614.8	71.17027	2359.648	2587240	1	0	1	2	5
3*1	358	125.2989	3947.246	750754.2	70.55484	2155.972	2417030	1	0	1	3	1
3*2	338	127.9911	4079.429	804033.2	71.41352	2298.903	1438188	1	0	1	3	2
3*3	366	126.7158	4020.281	939542.6	72.94950	2231.037	2532701	1	0	1	3	3
3*4	344	130.2616	3813.186	1073049.3	70.88187	2248.211	3383663	1	0	1	3	4
3*5	363	124.4160	4144.185	699090.4	70.59960	2228.464	1349448	1	0	1	3	5

Objname1

Paso 3:

sujeto a

$$\begin{cases} CV_{PO} \leq 0.03 \\ CV_{VENTAS} \leq 0.05 \\ CV_{VMP} \leq 0.05 \end{cases}$$

Creación del
objname2
con los
CV o errores

Objname2

```
> #####
> # Creacion del dataframe con las
> # CV deseados para las estimaciones
> #####
> DOM="DOM1"
> CV1=0.03
> CV2=0.05
> CV3=0.05
> domainvalue=1
> CV=data.frame(DOM,CV1,CV2,domainvalue)
> CV
      DOM   CV1   CV2 domainvalue
1 DOM1 0.03  0.05          1
```

Paso 4:

Solución

*dataframeName = Bethel(*objname1*,*objname2*,printa = TRUE)*

```
#####
# Busqueda de una solucion
# por el Algoritmo de Bethel
#####

solu=bethel(StratResumen,CV,printa=TRUE)
```

Resultado de la ejecución

```
> #####
> # Solucion alcanzada
> #####
> solu
[1] 18 20 22 19 18 25 24 21 23 21 24 23 26 24 25
attr(,"confr")
      STRATUM POPULATION BETHEL PROPORTIONAL EQUAL
[1,] "1*1"    "257"       "18"     "18"       "23"
[2,] "1*2"    "281"       "20"     "20"       "23"
[3,] "1*3"    "305"       "22"     "22"       "23"
[4,] "1*4"    "266"       "19"     "19"       "23"
[5,] "1*5"    "272"       "18"     "19"       "23"
[6,] "2*1"    "360"       "25"     "26"       "23"
[7,] "2*2"    "324"       "24"     "23"       "23"
[8,] "2*3"    "304"       "21"     "22"       "23"
[9,] "2*4"    "336"       "23"     "24"       "23"
[10,] "2*5"   "308"       "21"     "22"       "23"
[11,] "3*1"    "358"       "24"     "25"       "23"
[12,] "3*2"    "338"       "23"     "24"       "23"
[13,] "3*3"    "366"       "26"     "26"       "23"
[14,] "3*4"    "344"       "24"     "24"       "23"
[15,] "3*5"    "363"       "25"     "26"       "23"
[16,] "TOTAL"  "4782"     "333"    "340"      "345"
attr(,"outcv")
      TYPE DOMAIN/VAR. PLANNED CV ACTUAL CV SENSITIVITY 10%
[1,] "DOM1"  "1/V1"      "0.03"    "0.0296"  "65"
[2,] "DOM1"  "1/V2"      "0.05"    "0.0305"  "1"
```

Optimización con más de un Dominio

- ⊗ El problema planteado da respuesta a encontrar n_1, \dots, n_H /

$$\min f(n_1, \dots, n_H) = \min \left\{ C_0 + \sum_{h=1}^H c_h n_h \right\} \text{ sujeto a} \begin{cases} CV_{PO} \leq 0.03 \\ CV_{VENTAS} \leq 0.05 \\ CV_{VMP} \leq 0.05 \end{cases}$$

a nivel de U sobre un conjunto de G variables objetivos .

- ⊗ Si se tienen D dominios y se desea estimar con cierta precisión en algunos o en cada uno de ellos, y a nivel U , el planteo para el uso de "Bethel" se puede modificar para aceptar todas las restricciones.

Algoritmo de Bethel con Dominios

Sean D dominios, $U = \bigcup_{d=1}^D U_d$ $U_d \cap U_{d'} = \emptyset$

- ⊗ los dominios pueden ser estratos o no, unión de algunos y en general cruzan los estratos de diseño
- ⊗ para cada $u_k \in U$ se sabe por el Marco Muestral a que Dominio pertenece

se definen para cada variable $g = 1, \dots, G'$ $Y_{kgd}^* = \begin{cases} Y_{kg} & \text{si } k \in U_d \\ 0 & \text{si } k \notin U_d \end{cases}$

$$\min f(n_1, \dots, n_H) = \min \left\{ C_0 + \sum_{h=1}^H c_h n_h \right\}$$

sujeto a: $\begin{cases} CV_{Y_g}^2 \leq cv_g^2 & g = 1, \dots, G \\ CV_{Y_{gd}}^2 \leq cv_{gd}^2 & d = 1, \dots, D \quad y \quad g = 1, \dots, G' \end{cases}$

Ejemplo con 3 Dominios

PO, VENTAS y VMP con precisión del orden del 3%, 5% y 10% a nivel de U para PO y VENTAS a nivel de Región (REG) del orden del 5% y 10% en c/uno repectivamente

$$Y_1 = PO, \quad Y_2 = VENTAS, \quad Y_3 = VMP$$

$$Y_4 = \begin{cases} Y_1 & \text{si } \in REG = 1 \\ 0 & \end{cases}, \quad Y_5 = \begin{cases} Y_1 & \text{si } \in REG = 2 \\ 0 & \end{cases}, \quad Y_6 = \begin{cases} Y_1 & \text{si } \in REG = 3 \\ 0 & \end{cases}$$

$$Y_7 = \begin{cases} Y_2 & \text{si } \in REG = 1 \\ 0 & \end{cases}, \quad Y_8 = \begin{cases} Y_2 & \text{si } \in REG = 2 \\ 0 & \end{cases}, \quad Y_9 = \begin{cases} Y_2 & \text{si } \in REG = 3 \\ 0 & \end{cases}$$

$$\min f(n_1, \dots, n_{15}) = \min \left\{ C_0 + \sum_{h=1}^{15} c_h n_h \right\}$$

sujeto a :

$$CV_{Y_1} \leq 0.03, \quad CV_{Y_2} \leq 0.05, \quad CV_{Y_3} \leq 0.10$$

$$CV_{Y_d} \leq 0.05 \quad d=4, \dots, 6$$

$$CV_{Y_d} \leq 0.10 \quad d=7, \dots, 9$$

Definición por Dominios de la Y's y la variable que definen los Dominios

```
#####
# Busqueda de una solucion
# por el Algoritmo de Bethel
# n=? nh=? h=1,...,H
# CV precision para un Totales a NIVEL POBLACION
# CV precision para los Totales a NIVEL REGION
#####

# REG como Dominios

Marco$Y4=Marco$P0*ifelse(Marco$REG==1,1,0)
Marco$Y5=Marco$P0*ifelse(Marco$REG==2,1,0)
Marco$Y6=Marco$P0*ifelse(Marco$REG==3,1,0)
Marco$Y7=Marco$VENTAS*ifelse(Marco$REG==1,1,0)
Marco$Y8=Marco$VENTAS*ifelse(Marco$REG==2,1,0)
Marco$Y9=Marco$VENTAS*ifelse(Marco$REG==3,1,0)
```

El buildStrataDF() “objname1”

```
#####
# Creacion del dataframe con
# el resumen por estrato de los
# promedios y sd de cada variable Y
#####
```

```
StratDominio=buildStrataDF(Marco)
```

STRATO	N	M1	M2	M3	M4	M5	M6	M7	M8	M9	S1	S2	S3	S4	S5
1*1	257	126.3930	3915.428	1304743.8	126.3930	0.0000	0.0000	3915.428	0.000	0.000	72.20618	2388.266	4143005	72.20618	0.000
1*2	281	124.7794	4097.342	908154.7	124.7794	0.0000	0.0000	4097.342	0.000	0.000	73.15509	2384.987	1794076	73.15509	0.000
1*3	305	122.9967	3984.223	1006462.2	122.9967	0.0000	0.0000	3984.223	0.000	0.000	72.79126	2350.350	3171672	72.79126	0.000
1*4	266	126.5977	3790.173	1029411.3	126.5977	0.0000	0.0000	3790.173	0.000	0.000	72.03539	2355.482	3044145	72.03539	0.000
1*5	272	135.1801	3958.276	982938.1	135.1801	0.0000	0.0000	3958.276	0.000	0.000	67.90668	2322.701	1981940	67.90668	0.000
2*1	360	127.7972	3947.406	679325.8	0.0000	127.7972	0.0000	0.000	3947.406	0.000	70.42877	2336.979	1188628	0.00000	70.428
2*2	324	129.5988	3951.694	768263.6	0.0000	129.5988	0.0000	0.000	3951.694	0.000	74.87244	2342.474	1387096	0.00000	74.872
2*3	304	126.3026	4040.645	617368.3	0.0000	126.3026	0.0000	0.000	4040.645	0.000	72.21223	2315.137	1102061	0.00000	72.212
2*4	336	135.6577	3854.354	706093.5	0.0000	135.6577	0.0000	0.000	3854.354	0.000	69.74980	2205.686	1242219	0.00000	69.749
2*5	308	126.6916	4090.662	779614.8	0.0000	126.6916	0.0000	0.000	4090.662	0.000	71.17027	2359.648	2587240	0.00000	71.170
3*1	358	125.2989	3947.246	750754.2	0.0000	0.0000	125.2989	0.000	0.000	3947.246	70.55484	2155.972	2417030	0.00000	0.000
3*2	338	127.9911	4079.429	804033.2	0.0000	0.0000	127.9911	0.000	0.000	4079.429	71.41352	2298.903	1438188	0.00000	0.000
3*3	366	126.7158	4020.281	939542.6	0.0000	0.0000	126.7158	0.000	0.000	4020.281	72.94950	2231.037	2532701	0.00000	0.000
3*4	344	130.2616	3813.186	1073049.3	0.0000	0.0000	130.2616	0.000	0.000	3813.186	70.88187	2248.211	3383663	0.00000	0.000
3*5	363	124.4160	4144.185	699090.4	0.0000	0.0000	124.4160	0.000	0.000	4144.185	70.59960	2228.464	1349448	0.00000	0.000

Definición de los CV deseados “*objname2*” y Solución:

```
#####
# Creacion del dataframe con las
# CV deseados para las estimaciones
#####
```

```
DOM="DOM1"
# PO , VENTAS y VMP POBLACION
```

```
CV1=0.03
```

```
CV2=0.05
```

```
CV3=0.10
```

```
# PO REGION
```

```
CV4=0.05
```

```
CV5=0.05
```

```
CV6=0.05
```

```
# VENTAS REGION
```

```
CV7=0.10
```

```
CV8=0.10
```

```
CV9=0.10
```

```
domainvalue=1
```

```
CV=data.frame(DOM,CV1,CV2,CV3,CV4,CV5,CV6,CV7,CV8,CV9, domainvalue)
```

```
CV
```

A nivel de U

A nivel de Dominios

```
#####
# Busqueda de una solucion
# por el Algoritmo de Bethel
#####
```

```
solu=bethel(StratDominio,CV,printa=TRUE)
```

```
# Solucion alcanzada
solu
```

```
> # Solucion nh alcanzada
```

```
> solu
```

```
[1] 57 27 52 43 29 23 24 18 23 43 46 26 49 62 26  
attr(,"confr")
```

	STRATUM	POPULATION	BETHEL	PROPORTIONAL	EQUAL
[1,]	"1*1"	"257"	"57"	"30"	"37"
[2,]	"1*2"	"281"	"27"	"33"	"37"
[3,]	"1*3"	"305"	"52"	"35"	"37"
[4,]	"1*4"	"266"	"43"	"31"	"37"
[5,]	"1*5"	"272"	"29"	"32"	"37"
[6,]	"2*1"	"360"	"23"	"42"	"37"
[7,]	"2*2"	"324"	"24"	"38"	"37"
[8,]	"2*3"	"304"	"18"	"35"	"37"
[9,]	"2*4"	"336"	"23"	"39"	"37"
[10,]	"2*5"	"308"	"43"	"36"	"37"
[11,]	"3*1"	"358"	"46"	"42"	"37"
[12,]	"3*2"	"338"	"26"	"39"	"37"
[13,]	"3*3"	"366"	"49"	"42"	"37"
[14,]	"3*4"	"344"	"62"	"40"	"37"
[15,]	"3*5"	"363"	"26"	"42"	"37"
[16,]	"TOTAL"	"4782"	"548"	"556"	"555"

```
attr(,"outcv")
```

	TYPE	DOMAIN/VAR.	PLANNED CV	ACTUAL CV	SENSITIVITY	10%
[1,]	"DOM1"	"1/V1"	"0.03"	"0.0247"	"1"	"62"
[2,]	"DOM1"	"1/V2"	"0.05"	"0.0254"	"1"	"26"
[3,]	"DOM1"	"1/V3"	"0.1"	"0.0992"	"109"	"25"
[4,]	"DOM1"	"1/V4"	"0.05"	"0.038"	"1"	"548"
[5,]	"DOM1"	"1/V5"	"0.05"	"0.0488"	"1"	"333"
[6,]	"DOM1"	"1/V6"	"0.05"	"0.039"	"1"	
[7,]	"DOM1"	"1/V7"	"0.1"	"0.0403"	"1"	
[8,]	"DOM1"	"1/V8"	"0.1"	"0.0511"	"1"	
[9,]	"DOM1"	"1/V9"	"0.1"	"0.0389"	"1"	

Incremento del los tamaños
muestrales con/sin dominios
en el proceso de optimización

	BETHEL	BETHEL
[1,]	"57"	"18"
[2,]	"27"	"20"
[3,]	"52"	"22"
[4,]	"43"	"19"
[5,]	"29"	"18"
[6,]	"23"	"25"
[7,]	"24"	"24"
[8,]	"18"	"21"
[9,]	"23"	"23"
[10,]	"43"	"21"
[11,]	"46"	"24"
[12,]	"26"	"23"
[13,]	"49"	"26"

> # Solucion nhd por Dominio
> sum(solu[1:5])
[1] 208
> sum(solu[6:10])
[1] 131
> sum(solu[11:15])
[1] 209

ASPIRACIÓN MÁXIMA EN UNA ESTRATIFICACIÓN MULTIVARIADA

UNA ALTERNATIVA PARA BRINDAR SOLUCIONES A:

- A) ENCONTRAR LA MEJOR PARTICIÓN DE LA POBLACIÓN
- B) TAMAÑO DE MUESTRA Y ASIGNACIÓN
- C) MINIMIZACIÓN DE COSTOS

Planteo

- Dado:
- a) un conjunto de variables objetivos $Y_g, g = 1, \dots, G$
 - b) un conjunto de potenciales variables estratificadoras $X_m, m = 1, \dots, M$
 - c) restricciones cv_g sobre las estimaciones
 - d) una estructura de costos $C(\cdot)$ fijos sobre los estratos

"¿Se puede crear la "mejor" estratificación que minimice una función de Costos, fijar un tamaño de muestra asumiendo una asignación y precisiones dadas a nivel global y/o por dominios?"

- * La "mejor" estratificación equivale a encontrar $h=1, \dots, H=?$ que satisfaga los requisitos y/o restricciones de precisión

Propuesta

a) Construir una estratificación con las variables (X's):

PO (*continua*) REG(3) ACTIV(5)

Asumiendo una estructura de Costos fijos sobre los estratos

b) Buscar una precisión del orden de 0.05 en términos de CV para los totales de VENTAS y de VMP a nivel de U ,

$$\begin{aligned} cv(VENTAS) &\leq 0.05 \\ cv(VMP) &\leq 0.05 \end{aligned}$$

$$\begin{aligned} \min f(n_1, \dots, n_?) &= \min \left\{ C_0 + \sum_{h=1}^? c_h n_h \right\} \\ &\text{sujeto a } \begin{cases} CV_{VENTAS} \leq 0.05 \\ CV_{VMP} \leq 0.05 \end{cases} \end{aligned}$$

INTRODUCCIÓN A LOS ALGORITMOS GENÉTICOS (GA)

Notación y Elementos Principales
para el Uso de “SamplingStrata” en R

Representación de una Partición o Estratificación

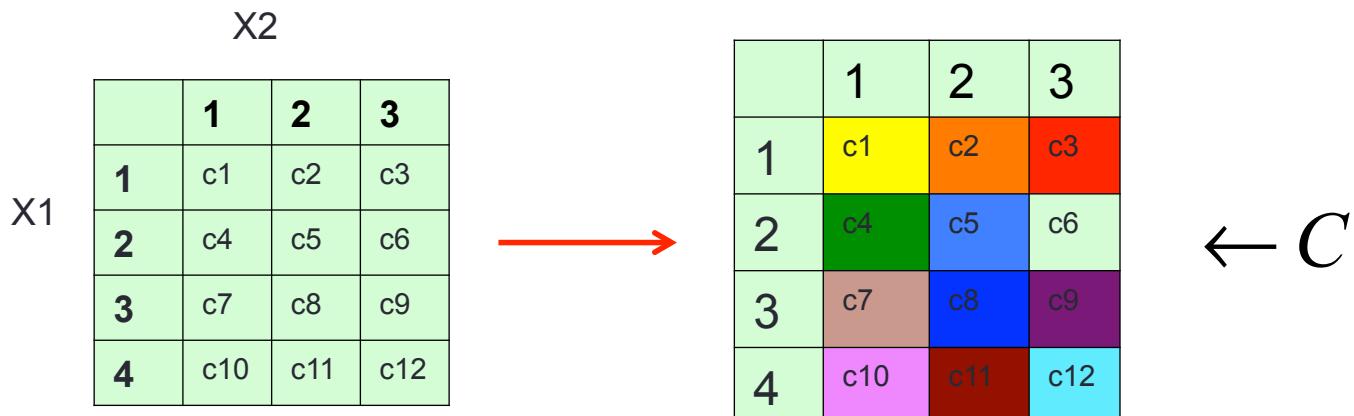
$$X_1 \quad d_1 = \{1, 2, 3, 4\} \quad X_2 \quad d_2 = \{1, 2, 3\} \quad K = 4 * 3 = 12$$

$$c_1 = celda_{11}, c_2 = celda_{12}, \dots, c_{12} = celda_{43}$$

$$C = PC^0 = \{c_1, \dots, c_{12}\}$$

C = Partición más fina de la U según X_1 y X_2

Es el conjunto de *microestratos* la mayor estratificación posible con X_1 y X_2



Otras Particiones o Estratificaciones Posibles

	1	2	3
1	c1	c2	c3
2	c4	c5	c6
3	c7	c8	c9
4	c10	c11	c12

$$H = 2$$

$$\left\{ \left\{ c_1, c_2, c_4, c_5 \right\}, \left\{ c_3, c_6, c_7, c_8, c_9, c_{10}, c_{11}, c_{12} \right\} \right\}$$

no es la única!

	1	2	3
1	c1	c2	c3
2	c4	c5	c6
3	c7	c8	c9
4	c10	c11	c12

$$H = 3$$

$$\left\{ \left\{ c_1, c_4, c_7, c_{10} \right\}, \left\{ c_2, c_3, c_5, c_6 \right\}, \left\{ c_8, c_9, c_{11}, c_{12} \right\} \right\}$$

	1	2	3
1	c1	c2	c3
2	c4	c5	c6
3	c7	c8	c9
4	c10	c11	c12

$$\{U\}$$

Espacio de todas las Particiones o el Universo de Estratificaciones

Sea $C = \{c_1, \dots, c_K\}$ y $\mathcal{P} = \{P_1, P_2, \dots, P_{B_K}\}$ el Espacio de todas las Particiones de C o el "Universo de Estratificaciones"

donde B es calculado por la fórmula de Bell:

$$B_K = \sum_{i=0}^{K-1} \binom{K-1}{i} B_i \quad (B_0 = 1)$$

B_K cuenta la cantidad de particiones de un conjunto de tamaño K

B_2	B_3	B_4	B_5	B_6	B_7	B_8	B_9	B_{10}	B_{11}	B_{12}
2	5	15	52	203	877	4140	21147	115975	678570	4213597

Tener en cuenta que dos variables estratificadoras con 2 categorías cada una y una tercera con 5 lleva a ... $2 \times 2 \times 5 = 20$ con lo cual:

$B_{20} = 51.724.158.235.372 \dots \text{ el Universo puede ser muy grande!}$

Búsqueda en el Universo de Estratificaciones

Si B_K es "manejable" se podría pasar por todas las particiones de U para minimizar $C(n_1, \dots, n_H)$.

Dada $P_l \in B_K$ caracterizada por H estratos tal que $U = \bigcup_{h=1}^H U_h$

implica resolver por el Algoritmo de Bethel

$$\begin{cases} \text{minimizar } C(n_1, \dots, n_H) \\ \text{con } C(n_1, \dots, n_H) = C_0 + \sum_{h=1}^H c_h n_h \\ \text{respetando } CV_{y_g} \leq cv_g \quad g=1, \dots, G \end{cases}$$

y elegir aquella partición que arrojó el menor $C(n_1, \dots, n_H)$

Pero si B_K es intratable hay que emplear algoritmos de búsqueda que atiendan la magnitud del espacio de soluciones y que identifique a la óptima o "alguna" que no esté lejos de ella.

Elementos Centrales de un Algoritmo Genético (GA)

*representación de una estratificación
a través del "genoma" de un individuo*

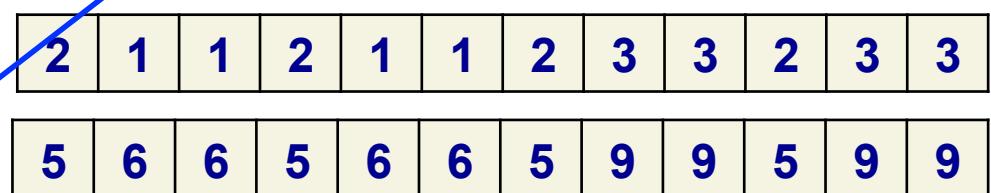
estratificación

	1	2	3
1	c1	c2	c3
2	c4	c5	c6
3	c7	c8	c9
4	c10	c11	c12

individuo

cromosoma

$$\left\{ \left\{ l_1, l_4, l_7, l_{10} \right\}, \left\{ l_2, l_3, l_5, l_6 \right\}, \left\{ l_8, l_9, l_{11}, l_{12} \right\} \right\}$$



*genomas equivalentes de un mismo
individuo o estratificación*

$$H = 3$$

*"distintas" etiquetas pueden
originar una misma estratificación*

Evolución en un GA

Universo de Estratificaciones
equivale a una Población de individuos
representados por sus "genomas" \Leftrightarrow *Soluciones posibles del problema*

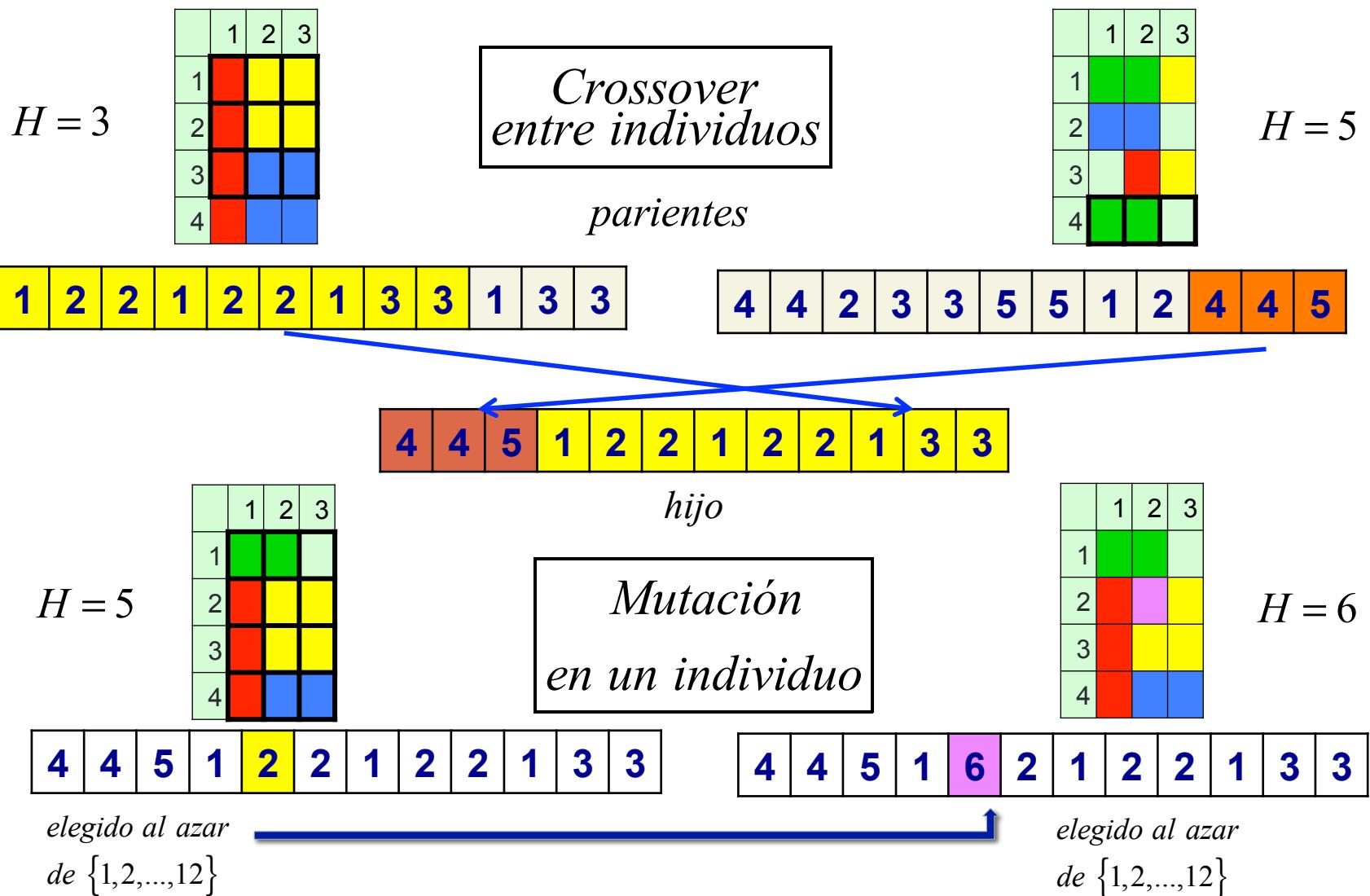
Generación = Población surgida de un proceso Evolutivo que implicó el nacimiento de nuevos individuos

Parientes = individuos de una generación anterior que son “padres” de una nueva Generación

Aptitud (fitness) = medida de calidad de un individuo en cada Generación

Objetivo : "A través de un proceso iterativo obtener nuevas generaciones a partir de una Población inicial, con aptitudes mayores que las precedentes"

Principales Operadores para obtener una Nueva Generación



De Generación en Generación

Generación J

2	1	1	3	3	3	1	1	1	2	1	2
4	3	3	1	1	2	1	1	1	2	4	2
5	5	4	4	2	2	2	3	3	1	1	1

Generación J+1

2	2	2	2	3	1	3	1	1	4	4	4
1	2	2	5	3	1	3	1	1	4	5	5
6	6	2	2	3	3	3	5	5	5	1	4

Crossover

4	4	4	1	1	2	1	2	2	1	1	2
1	1	1	2	2	2	2	3	3	3	1	1

Mutación

2	2	2	2	3	1	3	1	1	4	4	4
---	---	---	---	---	---	---	---	---	---	---	---

1	1	1	2	2	3	3	2	2	3	3	3
4	4	2	2	2	1	2	3	3	3	1	1

1	1	1	2	2	3	3	2	2	3	3	3
4	4	2	2	2	1	2	3	3	3	1	1

Elitismo

Principales Pasos en un GA para Estratificar

- **Paso 1:** Generar en forma **aleatoria** una Población Inicial de tamaño P (Un conjunto de Estratificaciones iniciales)
- **Paso 2:** Calcular la “**Aptitud**” de c/Padre o Estratificación para seleccionar las “más aptas” (los mejores C(n))
- **Paso 3:** **Seleccionar** pares de Padres o Estratificaciones. Aplicar “**crossover**” entre ellas y un % de “**mutación**” sobre las descendientes
- **Paso 4:** Pasar un % de la Generación anterior y a la nueva (la “**elite**” de una Generación, los mejores los C(n))
- **Paso 5:** Repetir Pasos 2 a 4 hasta un nro R de Generaciones predeterminadas en el inicio del algoritmo o bien la aptitud no cambia por un número determinado de Generaciones

Ajustando un GA para Estratificación

Paso 1: Generar P individuos (estratificaciones) aleatoriamente de \wp ,

Cada individuo se obtiene al seleccionar K cromosomas
 v_k al azar de $\{1, 2, \dots, V\}$, $V \leq K$

$$v = \boxed{v_1 \ v_2 \ . \ . \ . \ v_k \ . \ . \ . \ . \ . \ v_K}$$

$v = [v_1, \dots, v_K]$ define la partición $P(v)$ con D_i ($i = 1, 2, \dots, Q_{P(v)}$) estratos

$$P = 4, \ K = 12, \ V = 6 \quad v^p = [v_1, \dots, v_{12}] \quad p = 1, \dots, 4$$

v^1	1	2	2	1	2	2	1	3	3	1	4	4
v^2	2	6	1	1	3	3	5	4	5	1	2	4
v^3	2	4	2	5	1	3	1	6	1	6	3	1
v^4	6	6	2	1	2	2	1	4	3	4	5	5

No se desean estratificaciones con
 $H \geq 7$:
 v_k al azar de $\{1, 2, \dots, 6\}$, $6 \leq K$

Ajustando un GA para Estratificación

Paso 2: Resolver por Bethel para cada individuo (o *estratificación* o $P(v)$) para poder determinar la "aptitud"

$$\begin{cases} \text{minimizar } C(n_1, \dots, n_{Q_{P(v)}}) \\ \text{con } C(n_1, \dots, n_{Q_{P(v)}}) = C_0 + \sum_{h=1}^{Q_{P(v)}} c_h n_h \\ \text{respetando } CV_{y_g} \leq cv_g \quad g=1, \dots, G \end{cases}$$

$$aptitud_p = 1 - \frac{n_{p-BETHEL}}{\sum_{k=p}^P n_{p-BETHEL}} \quad n_{p-BETHEL} \text{ solución para el individuo } p$$

Ajustando un GA para Estratificación

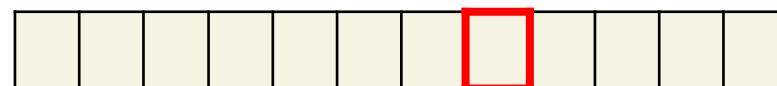
Paso 3: a) Seleccionar parientes proporcional a su $aptitud_p$ para conformar pares que engendrarán hijos.

$$aptitud_p = 1 - \frac{n_{p-BETHEL}}{\sum_{k=p}^P n_{p-BETHEL}}$$

$n_{p-BETHEL}$ solución para el individuo p

b) Seleccionar un cromosoma al azar entre los K posibles como punto de cruce para el *crossover* entre los parientes

$$k = 8$$



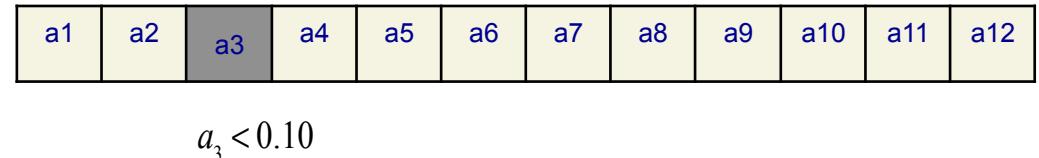
c) Fijar un % de *mutación* de *cromosomas* que pueden ser mutados al momento de generar un *hijo*.

Generar un $a_k \sim U(0,1)$, $k = 1, 2, \dots, K$ si $a_k < \%$ de *mutación* v_k cambia a otro valor seleccionado al azar de $\{1, \dots, V\}$

$$\% \text{mutación} = 10\%$$

$$12 \text{cromosomas} * 10\% = 1.2$$

≈ 1 cambio



Ajustando un GA para Estratificación

Paso 4: Pasar un % de la Población de los + aptos "la elite" a la próxima generación, %_elitismo.

Por ej:

%_elitismo=0.20 \equiv el 20% de los mejores $C(n)$ pasan a la próxima generación

O sea se preservan el 20% de las mejores estratificaciones de una generación a otra

Paso 5: En general el algoritmo termina cuando:

- (a) se alcanza un *máx* de iteraciones (*iter*)
o sea un número de generaciones previstas
- o (b) cuando la solución no mejora a través de las iteraciones
o sea cuando el $C(n)$ con mayor *amplitud*^p de una generación no cambia y se mantiene como la "mejor"^p a través de las generaciones

Observaciones sobre los Parámetros del Algoritmo

- ⊗ El tamaño de la población inicial (*pops*) en gral influye en la rapidez de convergencia del algoritmo,
- ⊗ Generar cada individuo lleva a conformar el vector v_k seleccionando K cromosomas al azar de $\{1, 2, \dots, V\}$, $V \leq K$. El parámetro **V** es importante porque define en principio el n^{ro} máximo de estratos "permitidos" en las estratificaciones (o sea el H)
- ⊗ La selección proporcional a la amplitud pasen con mayor chance los + aptos a la otra generación y que los - aptos tengan también probabilidad, para inducir "diversidad" y prevenir convergencias prematuras a "malas" soluciones (mínimos locales)

Observaciones sobre los Parámetros

- ⊗ el crossover entre los parientes permite que los hijos tengan muchas características de sus progenitores y así incrementar el promedio de la "aptitud" a través de la Evolución
- ⊗ el % de mutación es importante.
Valores grandes aseguran diferencias entre generaciones sucesivas y disminuye la posibilidad de estacionarse en un mínimo local de la función a optimizar.
Pero también la convergencia puede ser muy lenta.
Por el contrario, un valor muy bajo acelera la convergencia pero aumenta el riesgo de mínimos locales
- ⊗ el % de elitismo asegura que las buenas soluciones estén siempre presentes a través de las generaciones.

```
> library("SamplingStrata")
```

EL COMANDO “optimizeStrata”

- **optimizeStrata(**
strata=Objname1,
errors=Objname2,
iter=nro de Generaciones (20),
initialStrata=límite superior de estratos en cada solución (3000),
pops=dimensión de cada Población (50),
mut_chance=para cada individuo probabilidad de cambiar un
cromosoma (0.05),
elitism_rate=tasa de mejores soluciones que se deberá preservar (0.20),
minnumstr=nro mínimo de unidades asignado a un estrato (2)
addStrataFactor=probabilidad que en cada mutación se incremente el
nro de estratos con respecto al vigente (0.01),
writefiles=guarda dataframes y plots producidos por el algoritmo(FALSE))

Solución por “SamplingStrata”

a) Construir una estratificación con las variables (X's):

PO (*continua*) REG(3) ACTIV(5)

Asumiendo una estructura de Costos fijos sobre los estratos

b)

$$\min f(n_1, \dots, n_r) = \min \left\{ C_0 + \sum_{h=1}^r c_h n_h \right\} \quad \text{sujeto a} \quad \begin{cases} CV_{VENTAS} \leq 0.05 \\ CV_{VMP} \leq 0.05 \end{cases}$$

⊗ SamplingStrata requiere que las variables de recorrido continuo (PO) se reagrupen en categorías.

El package permite emplear el método *k – means* como ayuda para atender este requerimiento en las variables continuas.

> library(SamplingStrata)

Marco.Empresas.RData

```
> setwd("~/Downloads/Curso Cordoba 2015/Clases/")
> load("Marco.Empresas 2015.RData")
> library(SamplingStrata)
> head(Marco.Empresas)
```

	RAZON.SOCIAL	VMP	VENTAS	PO	TIPO	ACTIV	REG
1	VACACION SOCIEDAD ANONIMA	1592778	2271	91	3	3	1
2	BUDDENSIEG S A ARGENTINA I C I Y F	1400547	6084	118	2	4	1
3	MAMPER S A	682373	2320	8	3	2	1
4	PACUCA SA	5548441	158	194	2	5	1
5	AGROPECUARIA LA MARIA PILAR SOCIEDAD ANONIMA	2570007	257	200	2	5	1
6	GEN AVE S A	5047769	2053	191	3	5	1

```
> table(Marco.Empresas$REG)

 1   2   3
1381 1632 1769

> table(Marco.Empresas$TIPO)

 1   2   3
1608 1596 1578

> table(Marco.Empresas$ACTIV)

 1   2   3   4   5
975 943 975 946 943

> summary(Marco.Empresas$PO)

    Min.  1st Qu. Median      Mean 3rd Qu.      Max.
      5.0     65.0    128.0    127.8    190.0    250.0
```

Transformación y Definición de las Variables

Creación Variables
Objetivos Y's

```
#####
# Renombrar las variables Continuas (Y)
# que se desan estimar con precision
#####
Marco$Y1=Marco$VENTAS
Marco$Y2=Marco$VMP
```

Creación de un
Dominio (U)
“domainvalues”
Nivel Población

```
#####
# Sin Estimaciones a Nivel Dominio
# Cuidado tiene que ser NUMERIC
#####
Marco$domainvalue=1
```

Creación Variables X's
Discretización de las
Continuas

```
#####
# Conversion de las variables Continuas
# que definen la estratificacion (X)
#####
Marco$X1=as.factor(Marco$REG)
Marco$X2=as.factor(Marco$ACTIV)
Marco$X3=var.bin(Marco$PO,bins=8)
#####
max(Marco$TIPO)*max(Marco$ACTIV)*max(Marco$X3)
```

Creación del “*Objname1*”

strata=Objname1

Los Micro Estratos

buildStrataDF ()

```
##### K=120 Micro Estratos #####
#          X1*X2*X3
#          B120
#####
# N, Promedios y STD en cada
# Micro Estrato para pasar a la Optimizacion
#####
StratosMarco=buildStrataDF(Marco)
nrow(StratosMarco)
```

```
> max(Marco$REG)*max(Marco$ACTIV)*max(Marco$X3)
[1] 120
```



los “microestratos” son en total REG(3)*ACTIV(5)*PO(8)=120

Dataframe StratosMarco resultado de buildStrataDF()

STRATO	N	M1	M2	S1	S2	COST	CENS	DOM1	X1	X2	X3
1*1*1	38	3738.368	931662.1	2327.305	1442238.8	1	0	1	1	1	1
1*1*2	31	4300.484	769402.1	2476.687	980564.0	1	0	1	1	1	2
1*1*3	34	3627.059	882465.8	2224.601	1625804.7	1	0	1	1	1	3
1*1*4	30	4247.433	1196557.1	2327.137	1580791.0	1	0	1	1	1	4
1*1*5	29	4142.345	727630.6	2497.376	969089.8	1	0	1	1	1	5
1*1*6	30	3921.733	2284860.9	2491.608	5685625.7	1	0	1	1	1	6
1*1*7	34	3580.412	2204619.3	2336.700	8218730.3	1	0	1	1	1	7
1*1*8	31	3891.452	1469671.1	2336.036	4939985.7	1	0	1	1	1	8
1*2*1	37	4359.784	964689.1	2275.442	1504313.6	1	0	1	1	2	1
1*2*2	42	4580.905	1480447.0	2366.252	3400405.1	1	0	1	1	2	2
1*2*3	43	4313.395	1053923.1	2436.483	2028286.4	1	0	1	1	2	3
1*2*4	30	4507.667	734011.8	2403.548	782808.3	1	0	1	1	2	4
1*2*5	24	3647.875	908170.0	2263.340	934375.4	1	0	1	1	2	5
1*2*6	34	3339.382	688227.0	2409.129	952010.3	1	0	1	1	2	6
1*2*7	28	4212.286	490580.2	1931.412	401780.8	1	0	1	1	2	7
1*2*8	43	3672.209	722049.2	2488.951	1179347.4	1	0	1	1	2	8

Creación del “*Objname2*”

errors=*Objname2*

Se definen los CV de cada Variable Objetivo

```
> #####
> # Precisiones deseadas
> # sobre las variables Y
> # en cada dominio
> #####
>
> DOM="DOM1"
> CV1=0.05
> CV2=0.05
> domainvalue=1
> errores=data.frame(DOM,CV1,CV2,domainvalue)
> errores
   DOM  CV1  CV2 domainvalue
1 DOM1 0.05 0.05          1
```

dataframe “errores”

	DOM	CV1	CV2	domainvalue
1	DOM1	0.05	0.05	1

Optimización con optimizeStrata()

```
#####
#      Optimizacion      #
#####
solu=optimizeStrata(strata=StratosMarco,
                     errors=errores,
                     iter=200,
                     pops=30,
                     initialStrata=30,
                     addStrataFactor=0.00,
                     mut_chance=0.05,
                     elitism_rate=0.20,
                     minnumstr=5,
                     writeFiles=TRUE)
```

Tamaño del Universo de Estratificaciones
 B_{120} intratable para psar por todas ellas

Cada individuo de cada Generación
posee 120 cromosomas

- Se generan en forma aleatoria 30 individuos o estratificaciones (**pops=30**)
- Se estudian 200 Generaciones (**iter=200**) (se analizan $30*200=6000$ estratificaciones)
- Se permiten soluciones con un máximo de hasta un total de 30 Estratos (**initialStrata=30**)
- Se permite que hasta aprox 6 cromosomas de los 120 puedan cambiar por mutación (**mut_chance=0.05**, $120*0.05=6$)
- Los 6 mejores de cada generación se pasan a la nueva población (**elitism_rate=0.20**, $pops*0.20=6$)
- El nro de muestra por estrato debe ser =>5 (**minnumstr=5**)

Tamaño muestral alcanzado y H final

```
*** Domain : 1 1
Number of strata : 120
GA Settings
Population size      = 30
Number of Generations = 200
Elitism                = 6
Mutation Chance        = 0.05
```

```
*** Sample size : 1087
*** Number of strata : 67
-----
...written output to outstrata.txt
```

Opción *writeFiles=TRUE*

4 archivos

outstrata1.txt
plotdom1.pdf
results1.txt
solution1.txt

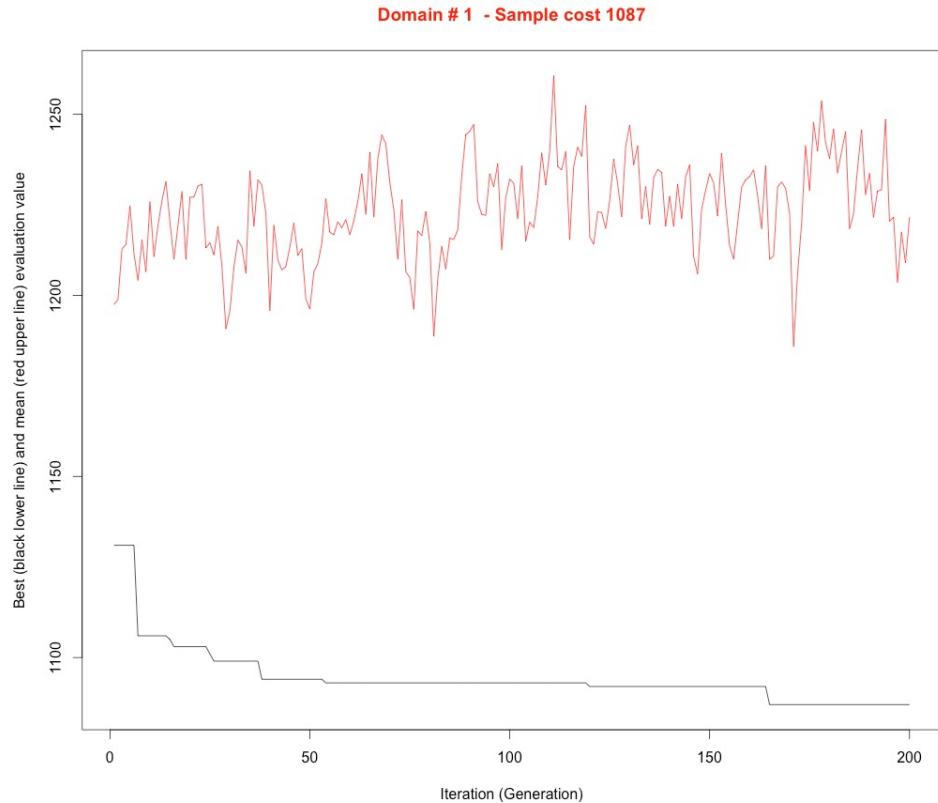
Convergencia del Algoritmo

A medida que evoluciona hacia la última generación (200) se muestra un gráfico de la convergencia de la solución (=tamaño de la muestra n).

- Con linea **roja** se visualiza el promedio (de los tamaños) de las 30 (pops=30) en c/Generación
- Con linea **negra** se detalla la mejor solución en c/iter o Generación

archivo

“plotdom1.pdf”



“result1.txt”

(historia de 6000 las soluciones)

```
Optimal stratification with Genetic Algorithm
-----
*** Parameters ***
-----
Domain: 1
Maximum number of strata: 120
Minimum number of units per stratum: 5
Take-all strata (TRUE/FALSE): FALSE
number of sampling strata : 120
Number of target variables: 2
Number of domains: 1
Number of GA iterations: 200
Dimension of GA population: 30
Mutation chance in GA generation: 0.05
Elitism rate in GA generation: 0.2
Chance to add strata to maximum: 0.01
Allocation with real numbers instead of integers: FALSE
*** Sample cost: 1087
*** Number of strata: 67
...written output to outstrata1.txt
```

“solution1.txt”

(nh final x Estrato)

59	
50	
22	
54	
46	
36	
14	
28	
31	
34	
24	
5	
16	
17	
18	
17	
34	
28	
38	
19	
18	
39	
15	
26	
4	
62	
17	
33	
10	
3	
11	

“outstat1.txt”

(resumen Estratos nuevos y nh (SOLUZ))

STRATO	M1	M2	S1	S2	N	DOM1	COST	CENS	SOLUZ
1	4049,3529	405740,1176	2088,7075	631623,1594	34	1	1	0	5
2	3777,4545	750256,6061	2296,8726	926397,3049	33	1	1	0	5
3	4451	705359,5161	2055,4017	966869,2932	31	1	1	0	5
4	3331,5077	771733,3538	2059,2954	869453,175	65	1	1	0	7
5	3999,2698	635167,1111	2129,7797	699068,7465	63	1	1	0	6
6	4683,1364	742647,2955	2341,6086	1396721,526	44	1	1	0	8
7	3671,0435	775244,3478	2070,1388	1159746,816	92	1	1	0	14
8	4069,5604	774033,9341	2190,4	1267963,124	91	1	1	0	15
9	4296,6477	516228,5909	2160,9845	484910,3587	88	1	1	0	6
10	3963,4928	1225167,797	2115,0617	4166068,688	69	1	1	0	36
11	3861,9684	659504,9304	2366,2684	813997,5177	158	1	1	0	16
12	4373,625	890173,625	1976,733	1706300,405	24	1	1	0	5
13	4737,7222	379212,3889	2505,8	355909,526	18	1	1	0	5
14	4295,1765	528334,9412	2267,8192	721037,2802	17	1	1	0	5
15	4058,1429	652852,5238	2460,1578	766546,0692	21	1	1	0	5
16	3578,875	975034,75	2289,9012	2414896,759	64	1	1	0	19
17	4299,3023	647642,1047	2097,4316	945674,2073	86	1	1	0	10
18	3550,3	2793769,85	2387,6977	9390165,574	40	1	1	0	40
19	3502,3043	1454985,565	2629,7351	3053667,803	23	1	1	0	9
20	4476	1562099	2274,5082	4815455,202	37	1	1	0	22
21	3588,2051	398918,3846	2278,676	445538,1318	39	1	1	0	5
22	4194,1266	441472,2278	2198,4192	422952,1065	79	1	1	0	5
23	4192,1852	1149689,519	2354,7553	1997690,147	54	1	1	0	14
24	3841,2222	658890,0988	2266,686	853041,8423	81	1	1	0	9
25	3073,2069	505575	2415,2053	775473,5303	00	1	1	0	0

Variables del objeto “solu” (67 estratos)

```
> # tamaño muestral  
> sum(ceiling(solu$aggr_strata$SOLUZ))  
[1] 1087  
> solu$aggr_strata  
   STRATO      M1        M2       S1        S2      N DOM1 COST CENS SOLUZ  
1     1 3924.500 720123.3 2442.355 1418012.3  52    1    1    0    10  
2     2 4030.296 641191.0 2114.112  572478.6  44    1    1    0     5  
3     3 4135.634 720842.8 2304.481 1675221.9 164    1    1    0    35  
4     4 3897.387 1012327.1 2197.446 1753216.8  80    1    1    0    18  
5     5 4015.586 824300.8 2314.416 1346154.9 198    1    1    0    34  
6     6 3868.818 1145268.7 2225.979 2948249.9  33    1    1    0    13  
7     7 3739.837 1394619.9 2096.723 5720459.2  49    1    1    0    36  
8     8 3504.206 1204324.3 2113.702 3205638.0  73    1    1    0    30  
9     9 3736.642 738077.1 2158.596 1732797.0  81    1    1    0    18  
10    10 3764.723 746298.5 2308.297 1982819.9  65    1    1    0    17  
11    11 3748.945 1101306.6 2220.069 2814944.6 127    1    1    0    45  
12    12 3698.891 605038.5 2359.309 733875.4  92    1    1    0     9  
13    13 3747.656 607757.1 2254.947 1777211.2 125    1    1    0    20
```

Total de “microestratos” (120) en cada uno de los 67 estratos

```
> table(solu$indices)  
  
 1  2  3  4  5  6  7  8  9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 29 30 31 32 33 34 35 36 37 38 39 40 41 42 43 44 45  
1  1  4  2  5  1  1  2  2  2  3  2  3  1  3  2  3  2  1  3  1  2  1  1  2  1  3  3  3  2  1  3  1  2  2  1  1  1  1  1  
46 47 48 49 50 51 52 53 54 55 56 57 58 59 60 61 62 63 64 65 66 68  
1  2  2  2  2  2  1  2  2  1  2  1  2  2  1  3  1  1  1  1  1  1
```

“solution1.txt” vincula microestrato con los estratos solución

```
# solution1.txt == solu$indices  
# copia en txt de los indices o LABELs que le corresponden a los  
# estratos viejos
```

```
Vinculos=read.table("solution1.txt")  
Vinculos
```

1	59
2	50
3	22
4	54
5	46
6	36
7	14
8	28
9	31
10	34
11	24
12	5
13	16
14	17
15	18
16	17
17	34
18	28
19	38
20	19
21	18
22	39
23	15
24	26
25	4
26	62
27	17
28	33
29	10
30	3
31	11
32	23

Solution1.txt

Vector de 120x2

por ejemplo

Estrato 17=microestratos {14,16, 27}

Actualización y Análisis de la Estratificación

- updateStrata(*objname1,solucion_por_optimizeStrata*)
- updateFrame(*dataframe_Marco,dataframe_nuevos_estratos*)

```
# Nuevos Estratos por  
# Reagrupamiento de los iniciales  
# Actualizcion de Estratos y del Marco
```

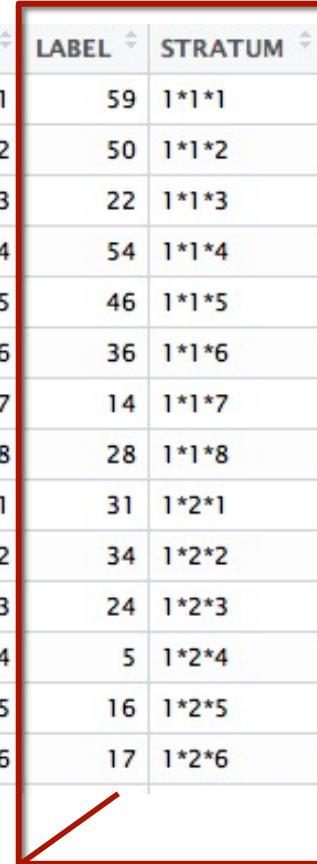
```
EstratosNuevos=updateStrata(StratosMarco,solu)
```

```
MarcoNuevo=updateFrame(Marco,EstratosNuevos)
```

dataframe EstratosNuevos

STRATO	N	M1	M2	S1	S2	COST	CENS	DOM1	X1	X2	X3	LABEL	STRATUM
1*1*1	38	3738.368	931662.1	2327.305	1442238.8	1	0	1	1	1	1	59	1*1*1
1*1*2	31	4300.484	769402.1	2476.687	980564.0	1	0	1	1	1	2	50	1*1*2
1*1*3	34	3627.059	882465.8	2224.601	1625804.7	1	0	1	1	1	3	22	1*1*3
1*1*4	30	4247.433	1196557.1	2327.137	1580791.0	1	0	1	1	1	4	54	1*1*4
1*1*5	29	4142.345	727630.6	2497.376	969089.8	1	0	1	1	1	5	46	1*1*5
1*1*6	30	3921.733	2284860.9	2491.608	5685625.7	1	0	1	1	1	6	36	1*1*6
1*1*7	34	3580.412	2204619.3	2336.700	8218730.3	1	0	1	1	1	7	14	1*1*7
1*1*8	31	3891.452	1469671.1	2336.036	4939985.7	1	0	1	1	1	8	28	1*1*8
1*2*1	37	4359.784	964689.1	2275.442	1504313.6	1	0	1	1	2	1	31	1*2*1
1*2*2	42	4580.905	1480447.0	2366.252	3400405.1	1	0	1	1	2	2	34	1*2*2
1*2*3	43	4313.395	1053923.1	2436.483	2028286.4	1	0	1	1	2	3	24	1*2*3
1*2*4	30	4507.667	734011.8	2403.548	782808.3	1	0	1	1	2	4	5	1*2*4
1*2*5	24	3647.875	908170.0	2263.340	934375.4	1	0	1	1	2	5	16	1*2*5
1*2*6	34	3339.382	688227.0	2409.129	952010.3	1	0	1	1	2	6	17	1*2*6

Nueva
Numeración



dataframe MarcoNuevo

STRATUM	RAZON SOCIAL	VMP	VENTAS	PO	TIPO	ACTIV	REG	Y1	Y2	X1	X2	X3	LABEL
1*1*1	ADROGUE SOCIEDAD ANONIMA	437787	4709	16	1	1	1	4709	437787	1	1	1	59
1*1*1	CULTIVOS LEONARDO QUADRINI SRL	439705	5780	28	2	1	1	5780	439705	1	1	1	59
1*1*1	PE#A MERCEDES JOSEFINA GAZTAMBIDE	54096	7417	13	3	1	1	7417	54096	1	1	1	59
1*1*1	GANADERA AGUAPEY SA	15702	1952	27	2	1	1	1952	15702	1	1	1	59
1*1*1	LAPHITZONDO RAUL MIGUEL	553128	4826	15	3	1	1	4826	553128	1	1	1	59
1*1*1	AGROPECUARIA LAS CATALINAS SA	977581	1052	10	2	1	1	1052	977581	1	1	1	59
1*1*1	LA JUANITA SOC EN COM POR ACCIONES	1079422	7970	27	1	1	1	7970	1079422	1	1	1	59
1*1*1	ZETONE Y SABBAG SA	5894957	5715	31	1	1	1	5715	5894957	1	1	1	59
1*1*1	ALBIOM SOCIEDAD ANONIMA	129871	1703	24	2	1	1	1703	129871	1	1	1	59
1*1*1	ESTANCIA SAN LUIS S C A	745033	2094	15	2	1	1	2094	745033	1	1	1	59
1*1*1	JUAN GERONIMO SOCIEDAD EN COMANDITA POR ACC...	365324	6419	10	3	1	1	6419	365324	1	1	1	59
1*1*1	AGROCOMERCIAL LATINOAMERICANA SA	396264	387	10	1	1	1	387	396264	1	1	1	59
1*1*1	CABALLERO MARIA CELIA LANUS	289756	2801	5	3	1	1	2801	289756	1	1	1	59
1*1*1	LUSTIG JULIA BRACERAS	850654	4139	14	3	1	1	4139	850654	1	1	1	59
1*1*1	EVRIS SRL	100705	7201	10	2	1	1	7201	100705	1	1	1	59

Selección de la MESA en c/Estrato

selectSample(dataframe,solu\$aggr _ strata)

```
> ### Seleccion de la muestra bajo MSA
> muestra=selectSample(MarcoNuevo,solu$aggr_strata)

*** Sample has been drawn successfully ***
1087 units have been selected from 67 strata

==> There have been 1 take-all strata
from which have been selected 34 units
> ## Resumen
> # Total Pob
> sum(muestra$WEIGHTS)
[1] 4782
> # Total de muestra por dominio
> table(muestra$DOMAINVALUE)

 1
1087
> # Total de unidades auto-representadas
> autorep=muestra[muestra$FPC==1,]
> sum(autorep$FPC)
[1] 34
```

La muestra

dataframe “muestra”

STRATO	STRATUM	RAZON.SOCIAL	VMP	VENTAS	PO	TIPO	ACTIV	REG	Y1	Y2	X1	X2	X3	LABEL	WEIGHTS	FPC
1	3*3*2	COOK MARIA ELINA MASCOTENA	227332	7177	52	2	3	3	7177	227332	3	3	2	1	5.200000	0.19230769
1	3*3*2	RAVERA ALBERTO JUAN	245709	4867	64	2	3	3	4867	245709	3	3	2	1	5.200000	0.19230769
1	3*3*2	CEBOL SUR SA	280824	5583	54	1	3	3	5583	280824	3	3	2	1	5.200000	0.19230769
1	3*3*2	BOSCH MARCELO	113968	3097	40	3	3	3	3097	113968	3	3	2	1	5.200000	0.19230769
1	3*3*2	AUTOMOTO SRL	435314	6415	59	3	3	3	6415	435314	3	3	2	1	5.200000	0.19230769
1	3*3*2	LA ESPERANZA S C A	228804	2299	58	2	3	3	2299	228804	3	3	2	1	5.200000	0.19230769
1	3*3*2	GAET S A	212364	5799	46	1	3	3	5799	212364	3	3	2	1	5.200000	0.19230769
1	3*3*2	CURARU SA AGROP COM	148022	1496	44	3	3	3	1496	148022	3	3	2	1	5.200000	0.19230769
1	3*3*2	FRUTAS GRIMALT SA	146972	5996	58	3	3	3	5996	146972	3	3	2	1	5.200000	0.19230769
1	3*3*2	ROSWHEY SOC ANONIMA	706566	6438	58	2	3	3	6438	706566	3	3	2	1	5.200000	0.19230769
10	1*4*5	IRUNDY SOCIEDAD ANONIMA AGRICOLA Y GANADER...	678963	1204	149	2	4	1	1204	678963	1	4	5	10	3.823529	0.26153846
10	3*4*2	EL SILENCIO S R L	691104	176	50	2	4	3	176	691104	3	4	2	10	3.823529	0.26153846
10	3*4*2	LA TORIBIA S A	265270	7463	47	1	4	3	7463	265270	3	4	2	10	3.823529	0.26153846
10	3*4*2	FUENTETOBA SAICIYA	141233	5548	65	3	4	3	5548	141233	3	4	2	10	3.823529	0.26153846
10	3*4*2	ECASIA	122162	7724	56	-	-	-	7724	122162	-	-	-	10	3.823529	0.26153846

Solución por Dominios con “SamplingStrata”

a) Construir una estratificación con las variables (X's):

PO (*continua*) REG(3) ACTIV(5)

Asumiendo una estructura de Costos fijos sobre los estratos

REG(3) Dominios de estimación

b) $\min f(n_1, \dots, n_?) = \min \left\{ C_0 + \sum_{h=1}^? c_h n_h \right\}$

sujeto a
$$\begin{cases} CV_{VENTAS_1} \leq 0.05 \\ CV_{VENTAS_2} \leq 0.07 \\ CV_{VENTAS_3} \leq 0.10 \\ CV_{VMP_1} \leq 0.10 \\ CV_{VMP_2} \leq 0.10 \\ CV_{VMPP_3} \leq 0.12 \end{cases}$$

Solución Cuando Hay Dominios

- ⊗ Cuando domainvalue $\neq 1$ el algoritmo trata a cada Dominio como si fuera una Población independiente y se ejecuta con los mismos parámetros y microestratos en cada uno de ellos y busca soluciones que satisfacen las restricciones.
- ⊗ Por lo tanto los tamaños muestrales se determinan según la estratificación como $n_h^* = \max \left\{ n_{1g}^*, \dots, n_{Hg}^* \right\}$
- ⊗ El tamaño muestral final $n^* = \sum_{h=1}^H n_h^*$

domainvalue=REG

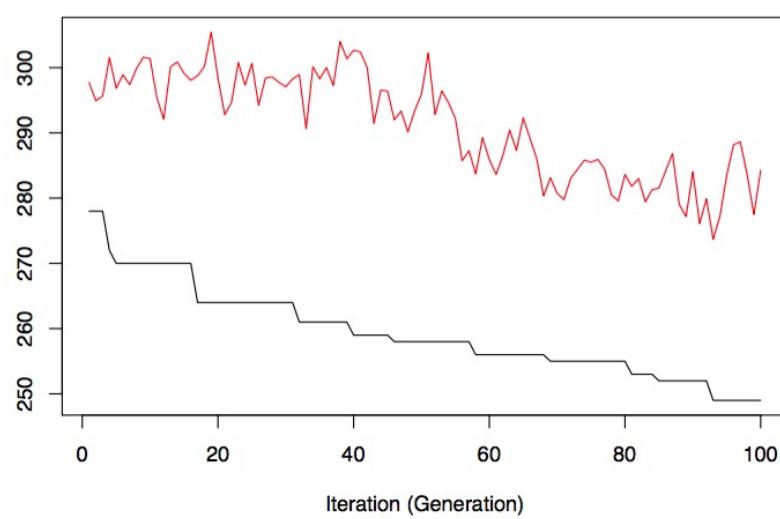
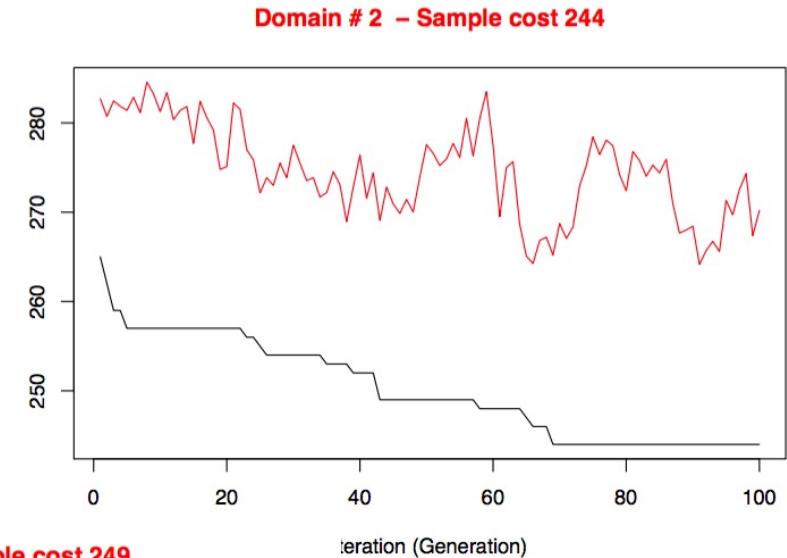
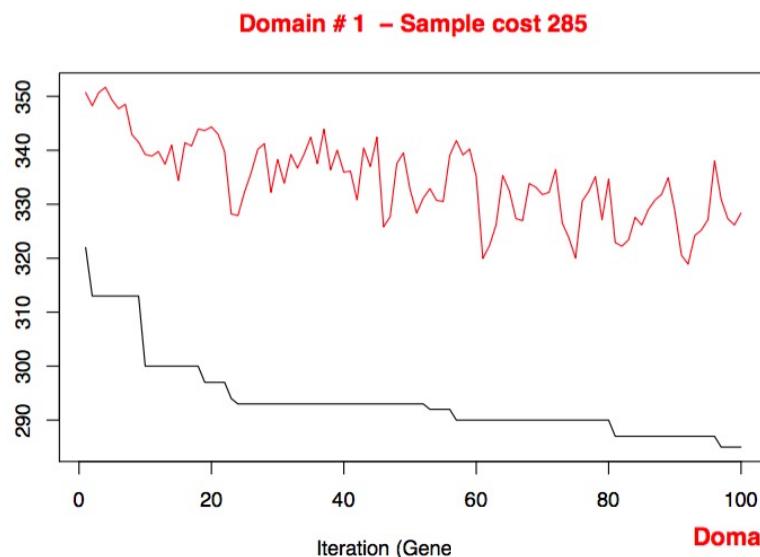
```
#####
#                               Solucion con Dominios
#####
Marco=Marco.Empresas
Marco$Y1=Marco$VENTAS
Marco$Y2=Marco$VMP
Marco$X1=as.factor(Marco$REG)
Marco$X2=as.factor(Marco$ACTIV)
Marco$X3=var.bin(Marco$PO,bins=8)

#####
# Los Dominios
#####
Marco$domainvalue=Marco$REG

#####
# Efectivos, Promedios y STD en cada
# Micro Estrato para pasar a la Optimizacion
#####
StratosMarco=buildStrataDF(Marco)
```

Convergencia de las Soluciones

(plotdom1.pdf , plotdom2.pdf , plotdom3.pdf)



Actualización del Marco y Análisis Posterior

```
> EstratosNuevos=updateStrata(StratosMarco,solu,writeFile=TRUE)
> MarcoNuevo=updateFrame(Marco,EstratosNuevos)
> attach(EstratosNuevos)
> Ordenado=EstratosNuevos[order(DOM1,LABEL),]
> ComoAgrego=read.delim("strata_aggregation.txt")
> # Muestra por Estrato
> tapply(solu$aggr_strata$SOLUZ,solu$aggr_strata$DOM1,sum)
  1   2   3
285 244 249
:
```

```
evalSolution(dataframe=name,
             solu$aggr_strata,
             nsample=value,
             writeFiles=TRUE)
```

```
# Evaluacion de la solucion sobre un conjunto de muestras

evalSolution(MarcoNuevo,solu$aggr_strata,nsampl=50,writeFiles=TRUE)

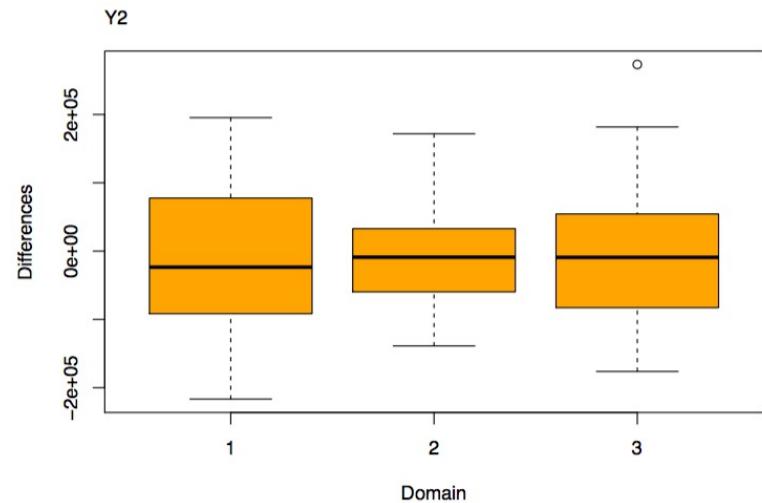
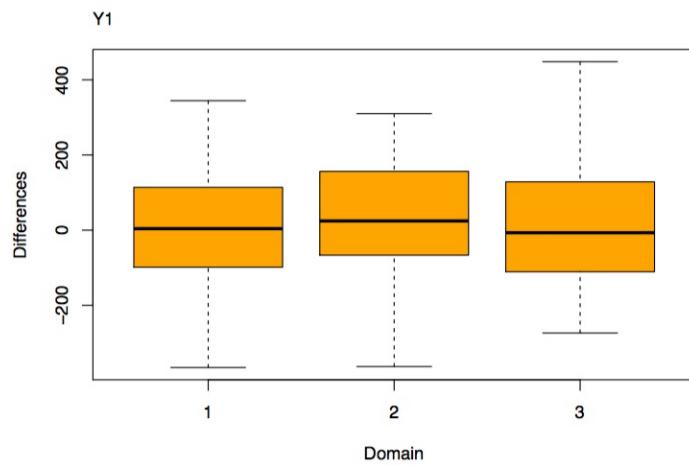
CV_esperados=read.csv("expected_cv.csv")

CV_esperados
```

```
> CV_esperados
```

	CV1	CV2	dom	DOM	CV1	CV2	domainvalue
1	0.04031248	0.10040728	DOM1	DOM1	0.05	0.10	1
2	0.03758664	0.09697691	DOM2	DOM1	0.07	0.10	2
3	0.04241324	0.12096106	DOM3	DOM1	0.10	0.12	3

Archivo differences.pdf (creado por evalSolution)



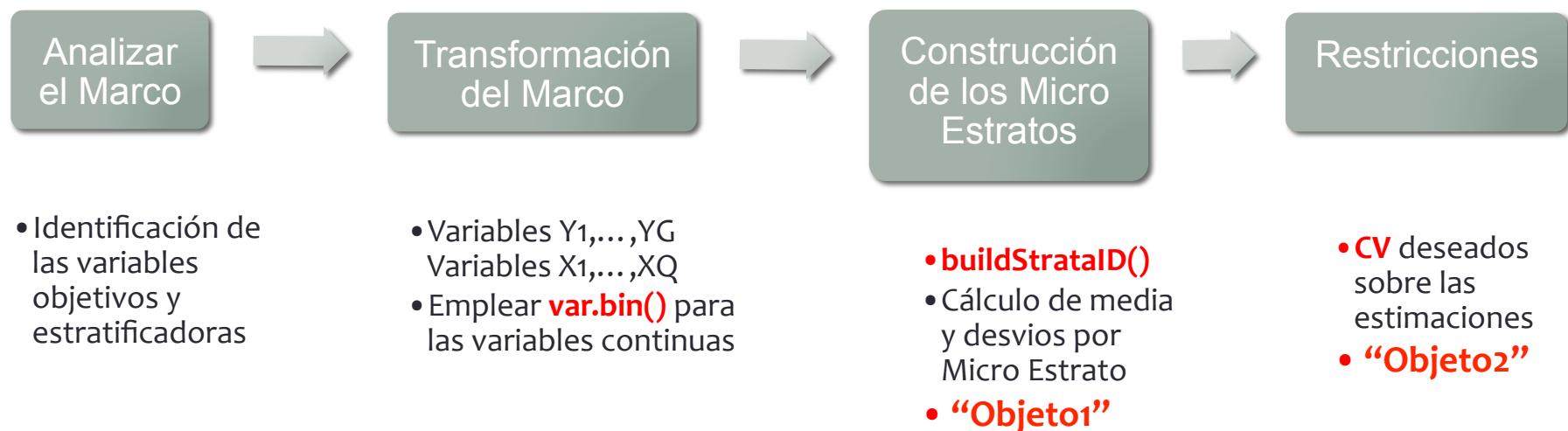
Selección de la Muestra de tamaño 778

```
> ### Seleccion de la muestra bajo MSA
> muestra=selectSample(MarcoNuevo,solu$aggr_strata)

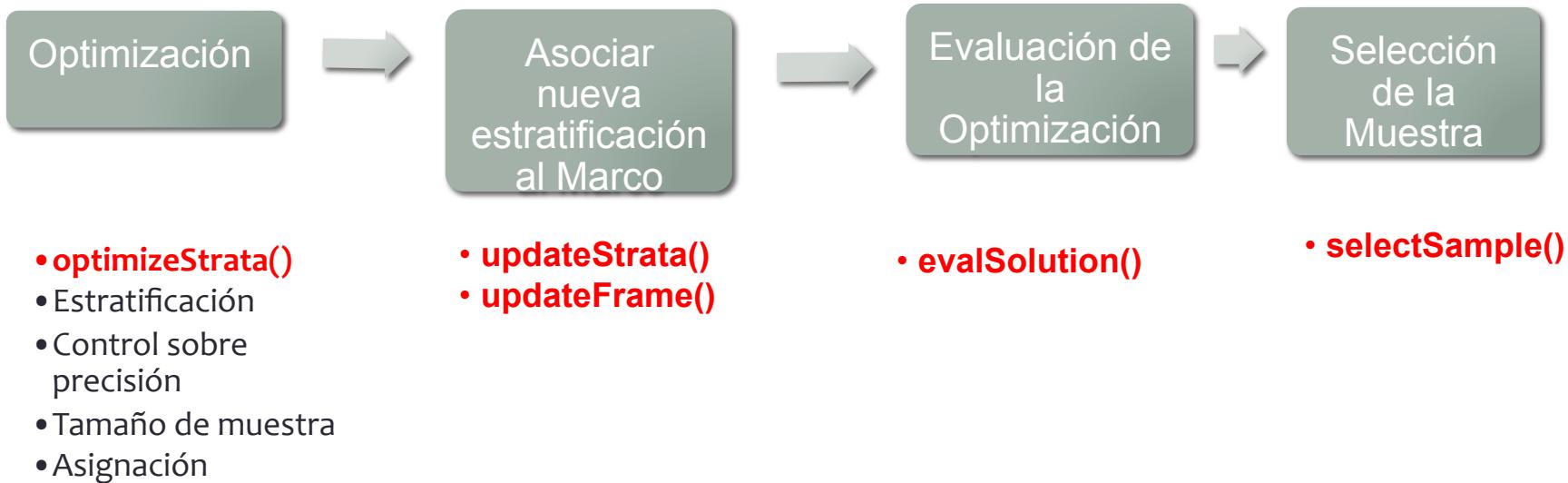
*** Sample has been drawn successfully ***
778 units have been selected from 54 strata
> ## Resumen
> # Total Pob
> sum(muestra$WEIGHTS)
[1] 4782
> # Total Pob por dominio
> tapply(muestra$WEIGHTS,muestra$DOMAINVALUE,sum)
    1   2   3
1381 1632 1769
> # Total de muestra por dominio
> table(muestra$DOMAINVALUE)

    1   2   3
285 244 249
> # Total de unidades auto-representadas
> autorep=muestra[muestra$FPC==1,]
> sum(autorep$FPC)
[1] 0
> # Total de autorepresentadas por dominio
> tapply(autorep$WEIGHTS,autorep$DOMAINVALUE,sum)
    1   2   3
NA NA NA
> |
```

Hoja de ruta antes de la Optimización



Hoja de ruta post Optimización



FIN
