Revista Colombiana de Estadística
Volumen 28 N° 1. pp. 39 a 62. Junio 2005

Estrategia de muestreo para la estimación de la tasa de favoritismo en la elección presidencial

Leonardo Bautista Sierra*

Resumen

Se fijan los objetivos y se definen los criterios metodológicos de una Encuesta Nacional de Favoritismo en Elecciones Presidenciales. Utilizando el hecho de que el candidato H. Serpa se presentó como candidato en 1998 y en 2002 se utilizan los resultados electorales de 1998 para generar, en combinación con datos censales de 1993, una base de datos, con la que se construye la estrategia muestral para estimación de resultados del 2002. Se llega a un diseño con cuatro estratos de municipios. Uno de inclusión forzosa con las más importantes ciudades del país, un segundo estrato de ciudades intermedias, el tercer estrato con 610 municipios y un último estrato de municipios muy pequeños y de difícil acceso. A modo de prueba, se realiza un ejercicio computacional de hacer 500 estimaciones del favoritismo de los candidatos en 2002 con 500 muestras diferentes seleccionadas de acuerdo al plan propuesto. En el 96 % de los casos se habría acertado dando a Uribe como ganador absoluto en la primera vuelta. Se alcanzó una confiabilidad del 94,8 % y una precisión equivalente a un c.v.e igual a 3,9 %. Finalmente, se aplica la metodología sugerida para producir una muestra para 2006 previendo la presentación de A. Uribe como candidato a la presidencia. Se concluye con una muestra de 85 municipios, 6.400 manzanas para empadronar y 15.800 personas a entrevistar.

***Palabras Claves**: Muestra electoral, muestra probabilística, confiabilidad, precisión, estrategia de muestreo, estratificación electoral.*

1. Conceptos básicos

1.1. Introducción

La ley colombiana vigente a finales del siglo XX e inicio del XXI establece que el candidato que en el comicio obtenga el favor de al menos la mitad más uno de los votantes se convierte en el presidente electo para gobernar al país

*Profesor asociado. Departamento de Estadística. Universidad Nacional de Colombia. Sede Bogotá. E-mail: jlbautistas@unal.edu.co; lbautista@cable.net.co

durante un período de cuatro años. Si ningún candidato alcanza tal magnitud de favoritismo, habrá un nuevo comicio electoral, denominado segunda vuelta. En él, la población decide entre solo dos candidatos, aquellos que en la primera vuelta obtuvieron la mayoría absoluta. En las elecciones para Presidencia de 1994 (RNEC 1994) se realizaron dos vueltas, y los candidatos de la segunda vuelta fueron el ganador Ernesto Samper y el perdedor Andrés Pastrana. Cuatro años después, Andrés Pastrana se presentaría otra vez como candidato y tendría que someterse de nuevo a un proceso de dos vueltas (RNEC 1998), pero en esa ocasión ganaría la Presidencia frente al candidato Horacio Serpa. Cuatro años más tarde (RNEC 2002), es Serpa quien vuelve a presentarse como candidato y pierde en la primera vuelta frente al candidato Álvaro Uribe.

El porcentaje de votos por cada candidato, en cada uno de los municipios del país, cambió a lo largo de los ocho años, en el sentido de que el candidato perdedor en 1994 fue ganador en 1998, y el perdedor en la segunda vuelta de 1998, perdió en 2002 en la primera vuelta. Sin embargo, las diferencias entre municipios se mantienen, respetando ancestrales patrones de comportamiento electoral (Bautista & Pacheco 1989). Así por ejemplo, el 88.4 % de los municipios en los que Pastrana ganó en la segunda vuelta de 1994 fue también ganador en la segunda vuelta de 1998, el 75.6 % de los municipios en los que Serpa perdió en 1998, volvió a perder de forma contundente, menos del 20 % de favoritismo, en 2002 (Véase anexo 1.). En los comicios 1994, 1998 y 2002 el candidato del Partido Conservador Andrés Pastrana y el candidato derechista Álvaro Uribe dominaron en aquellos municipios y capitales de departamento, que históricamente han favorecido con su voto a los candidatos del partido Conservador. Se trata de las llamadas capitales "de clima frío" Manizales, Tunja, Pasto, otras tradicionalmente conservadoras como Medellín y municipios de corte más rural que urbano. Por el contrario, las poblaciones de "clima cálido" y en particular las de las dos costas Buenaventura, Cartagena, Barranquilla, Montería, Turbo, entre otras, le son regularmente favorables a los candidatos del partido Liberal. La propuesta metodológica que aquí se presenta aprovecha este comportamiento sistemático, para construir una estrategia muestral, confiable, precisa y económicamente viable para las encuestas de opinión electoral en comicios presidenciales.

1.2. Objetivo de una encuesta nacional de favoritismo en elecciones presidenciales

El objetivo de una encuesta nacional de favoritismo en elecciones presidenciales (ENFEP) es estimar la tasa de favoritismo que obtendrían determinados candidatos, si el comicio electoral fuera "hoy". Se trata de estudiar en forma anticipada el proceso que se da el día de elecciones. Dicho proceso se describe, desde el punto de vista de la teoría del muestreo y de manera simplificada, de la siguiente forma:

Llamando U al universo de personas mayores de 18 años del país con plenos derechos civiles, e indagados uno a uno en forma independiente y voluntaria, se establecen dos variables para cada persona, zk que establece si la persona es participante o abstencionista, y la variable y_k que señala si la persona vota por el

candidato particular $\boldsymbol{Y}$ o no lo hace, bien porque no vota o porque apoya a otro candidato.

$$\begin{aligned} z_k &= 0 \quad \text{si la } k\text{-ésima persona es abstencionista,} \\ z_k &= 1 \quad \text{si la } k\text{-ésima persona vota,} \\ y_k &= 0 \quad \text{si la } k\text{-ésima persona es abstencionista o} \\ &\qquad \text{participando no apoya al candidato } \boldsymbol{Y} \\ y_k &= 1 \quad \text{si la } k\text{-ésima persona vota y lo hace por el candidato } \boldsymbol{Y} \end{aligned} \tag{1}$$

El resultado electoral, que se divulga al concluir el día de elecciones, es la tasa de favoritismo para el candidato $\boldsymbol{Y}$, establecida como el cociente entre la cantidad de votos por el candidato (N_y) sobre la cantidad de votos válidos en el comicio (N_z).

$$R_y = \frac{\sum_U y_k}{\sum_U z_k} = \frac{N_y}{N_z} \tag{2}$$

Se trata, en términos técnicos, de una *tasa* y no de una *proporción*. La sutil, pero determinante diferencia entre estos dos conceptos es que las *tasas* se establecen con base en denominadores desconocidos y aleatorios, mientras que las *proporciones* se fundamentan en denominador constante y conocido de antemano (Bautista 1998). Para la ENFEP el denominador es la cantidad de votos entregados por la población. Es decir, es la cantidad de participantes en el comicio. La abstención electoral en Colombia es alta y variable entre municipios y sectores poblacionales, lo que convierte a la cantidad de participantes en cifra aleatoria y variable[1].

1.3. Metodología de una encuesta nacional de favoritismo en elecciones presidenciales

El método que utiliza el estadístico, y en particular el muestrista, para conformar su plan de estimación responde a tres preguntas básicas: Qué se va indagar, a quiénes, y cuál es la calidad del resultado que se entrega. Para predecir el resultado de elecciones, unas semanas antes del comicio, se realiza una entrevista directa a personas mayores de 18 años, de una parte muy particular del universo, en la que básicamente se plantean dos preguntas: 1.- ¿Votaría Usted, si las elecciones fueran hoy? 2. Si no, muchas gracias. Si sí, ¿Por quién votaría?

La forma como se plantean las preguntas, y posteriormente, la forma como se codifican y procesan las respuestas conducen a muy diferentes resultados de la estimación. La muestra o subconjunto de personas que dan su respuesta en la ENFEP, y cuya opinión es utilizada para estimar la opinión de los ciudadanos del país, debe ser tomada, siguiendo estrictas normas técnicas, para configurar lo que denomina una muestra probabilística, que dista mucho de ser sinónimo de

[1] El censo nacional de población de septiembre de 1993 arrojó una población de 19'109.852 personas mayores de 18 años. Nueve meses después, en las elecciones para Presidencia, la cantidad de votos válidos fue de 7'384.845, lo que arroja una abstención del 61.3 %. Cuatro años después, en 1998, la cantidad de votos válidos pasó de 10'626.000 votos en la primera vuelta a 12'180.000 en la segunda.

una muestra al azar. Por último, aunque se cumplan los criterios técnicos para el tratamiento de preguntas y respuestas, y se establezcan muestras que respetan el rigor de la teoría de muestreo, algunas decisiones técnicas del proceso de encuesta pueden conducir a resultados de poca confiabilidad o de muy corta precisión.

1.3.1. Las preguntas que se plantean y la codificación de las respuestas

Respecto a las preguntas que se plantean en la ENFEP, se trata aquí de aquellas que además de constituir una fotografía, modifican lo que se suele denominar la opinión pública. Un estudio que realiza un candidato y cuyos resultados son utilizados, sólo por sus coordinadores de campaña para orientar sus acciones, puede contener muy diferentes preguntas y formas de preguntar. Por ejemplo ¿Quién cree que ganaría, si las elecciones fueran hoy? ó ¿Si las elecciones fueran hoy, cuál candidato le gustaría que ganara?. Para una ENFEP destinada a la opinión pública, la pregunta o las preguntas básicas deberían referirse sin ambigüedad al interrogante, que el ciudadano del común cree que se le está respondiendo con los resultados de la encuesta (Gawiser & Witt 2002). Ese interrogante es:

1.- ¿Votaría Usted, si las elecciones fueran hoy?

2. Si la respuesta es "no", muchas gracias[2].

Si la respuesta es "sí", ¿por quien votaría?

El segundo aspecto a considerar es el relativo a la interpretación de la respuesta. La respuesta a si votaría hoy, puede tener seis opciones: *no sabe, no desea responder, seguramente no, probablemente no, seguramente sí y probablemente sí.* Desde el punto de vista de la calidad final del proceso de estimación, lo conservador es reducir el tamaño del denominador, considerando como respuestas "No" las primeras cuatro opciones.

Desde el punto de vista de cómo preguntar, se puede optar por la entrevista cara a cara, en la que el entrevistador enseña al entrevistado un símil del tarjetón electoral, al momento que formula la pregunta sobre preferencia (Biemer, Folsom, Kulka, Lesler, Shah & Weeks 2003). Este procedimiento costoso puede remplazarse por la entrevista telefónica, método más barato, pero basado en la memoria que tiene la población sobre los candidatos que participan en el comicio. El recuerdo espontáneo puede existir durante las últimas semanas de un proceso electoral[3] y en los casos de segunda vuelta, pero se puede llegar a resultados con distorsiones graves, si se supone equivocadamente, que la población tiene buena memoria sobre los candidatos y sus programas, en los momentos iniciales del debate.

[2]En ocasiones se pregunta la razón de la abstención, si ha votado en comicios anteriores, y otros aspectos relacionados con el tema de la abstención. Estas preguntas encarecen el estudio y se apartan del objetivo de la ENFEP

[3]El recuerdo de los candidatos participantes en el debate puede no existir en la población, incluso el mismo día de elecciones, en procesos electorales de menor importancia como la de dignatarios locales o regionales

1.3.2. Muestra probabilística

Es un error estadístico utilizar, para una encuesta cualquiera y en particular para una ENFEP, el método de entrevistar al azar a algunas personas a la salida de un supermercado, a algunos conductores de los que se detienen ante un semáforo en rojo, o a quien fortuitamente responde al teléfono. Éstos o similares procedimientos conforman muestras al azar, pero no necesariamente probabilísticas.

Una muestra, para ser considerada probabilística, debe cumplir (Särndal, Swensson & Wretman 2003): *Los elementos son seleccionados de un marco de muestreo, siguiendo un algoritmo que corresponde a probabilidades positivas y conocidas antes de la selección.* Aunque la probabilidad de selección de un número telefónico sea positiva y conocida, no lo es la probabilidad de que quien responda sea determinada persona del hogar. Tampoco se conoce la probabilidad de que un determinado conductor, el día de la entrevista escoja la ruta A o B, o que una persona decida visitar uno u otro supermercado.

Un Marco de Muestreo es un dispositivo (lista, mapa, directorio, etc.) que permite *identificar* y *ubicar* a cada uno de los elementos del universo de estudio. Para el caso de la ENFEP se necesita un marco de las personas adultas aptas para la entrevista. Este dispositivo se llama padrón y permite conocer el nombre y la dirección de cada uno de los residentes de una vecindad. Él existe en algunos países, pero no en Colombia. Su ausencia exige, desde el punto de vista del muestreo, que la selección de la muestra se realice en dos o más etapas. Es decir, seleccionar grandes conglomerados, como por ejemplo municipios; y dentro de los municipios seleccionados escoger algunas manzanas, realizar el empadronamiento de las personas mayores de 18 años de esas manzanas y, de ese padrón escoger aleatoriamente los nombres, con sus respectivas direcciones, de las personas que responderán a la entrevista de favoritismo electoral.

El proceso de muestreo en varias etapas consiste en establecer una partición[4] del universo de votantes. Los subconjuntos que forman la partición se denominan, para el muestreo, *Conglomerados primarios de muestreo - CPMs.* Se selecciona una muestra probabilística de esos conglomerados y se aplica un nuevo plan de muestra[5] al interior de cada conglomerado escogido en la primera etapa. Para una selección directa de elementos, es decir en el caso de la encuesta electoral de personas mayores de 18 años, se necesita el padrón a nivel de ese conglomerado, el que, o bien se construye o se aplica de nuevo un diseño en etapas. Para conformar una segunda etapa de muestreo en cada municipio, se realiza una partición, que para el caso puede construirse a partir de barrios, comunas, sectores cartográficos o manzanas. Las partes que conforman esta segunda partición se denominan *Conglomerados secundarios de muestreo - CSMs.* Se efectúa entonces una selección aleatoria de CSMs, con la mismas características dadas para la selección de CPMs. Si todavía se trata de segmentos geográficos muy grandes para hacer un levantamiento censal, se puede, sólo en los casos necesarios, proponer una terce-

[4]Conjunto de subconjuntos del universo que cumplen: no ser vacías, no traslaparse y su unión reconstruye el universo

[5]Cada proceso de selección debe respetar los principios de independencia e invarianza muestral.

ra etapa en la que se crean los *Conglomerados terciarios de muestreo - CTMs*, y así sucesivamente.

La literatura en lengua inglesa utiliza el término *listing* para designar el proceso por el que, se pasa vivienda por vivienda, en una manzana o en un grupo de manzanas, escribiendo los nombres de las personas mayores de 18 años, para realizar después, basándose en esa lista o padrón, la selección probabilística de los nombres de las personas que responderán a la entrevista. Dicho proceso se denomina aquí *empadronamiento*.

El proceso metodológico de selección de muestra descrito, se resume entonces en los pasos siguientes:

1. Realizar varias etapas de división, selección muestral, subdivisión, selección muestral, hasta llegar a una muestra de pedazos de manzanas, de manzanas o de grupos de manzanas.
2. Realizar el empadronamiento, es decir levantar en esos pedazos, manzanas o grupos de manzanas la lista completa de identificación y ubicación de las personas mayores de 18 años aptas para votar[6].
3. Establecer la muestra de personas, con nombre y ubicación precisas.
4. Realizar la entrevista, única y expresamente, a las personas seleccionadas en la muestra.

La aplicación de estos cuatro pasos sin vigilar cuidadosamente todos los requerimientos técnicos que ellos exigen, conduce a sesgos que, como se explica a continuación, afectan la confiabilidad y pueden hacer inútiles los resultados del estudio.

1.3.3. Estrategia muestral y sus criterios de calidad

El trabajo del muestrista consiste en escoger un modo de seleccionar muestras, *diseño de muestra*, y una fórmula de procesamiento de los datos observados, *estimador*, a fin de producir, al menor costo posible, un intervalo de amplitud pequeña, que con alta probabilidad contenga "la verdad', es decir el verdadero porcentaje que se está estimando. A la combinación de diseño y estimador, $[p(\cdot), \hat{R}(\cdot)]$ se le llama *la estrategia de muestreo* y al intervalo que se produce se le denomina *Intervalo de confianza*. Con esta terminología, el objetivo del muestrista es entonces escoger una estrategia muestral a fin de producir, a bajo costo, un intervalo de confianza, tal que la probabilidad de que la "verdad" esté cubierta por él, sea muy alta, es decir, tal que:

$$P\big(\, R_y \in \big[\hat{R}_y - z_{1-\frac{\alpha}{2}}\sqrt{V_p(\hat{R}_y)}\,,\; \hat{R}_y + z_{1-\frac{\alpha}{2}}\sqrt{V_p(\hat{R}_y)}\,\big]\,\big) = P_c \qquad (3)$$

[6]Se suele preguntar además por el sexo, la edad y el número telefónico (Bautista 2000). Se pregunta el sexo para evitar situaciones incómodas a los entrevistadores puesto que hay nombres, de los que no se sabe si se trata de mujeres o de hombres. La edad para diferenciar, por ejemplo, padres e hijos homónimos; y el número del teléfono para concertar citas, solicitar aclaraciones o para realizar los operativos de supervisión de campo.

Obviamente, sin necesidad de recurrir al muestreo estadístico, se sabe que el porcentaje de favoritismo de un determinado candidato está con probabilidad uno, entre el cero y el cien por ciento. De tal intervalo se dice que es confiable porque tiene probabilidad uno de acierto, $P_c = 1$, pero que es impreciso porque aporta un conocimiento inútil. Dependiendo del diseño y del estimador, es decir de la estrategia muestral que se aplique, la probabilidad P_c puede hacerse grande o pequeña. También la longitud del intervalo, determinada por la varianza del estimador $V_p(\hat{R})$ depende de la estrategia muestral. A la probabilidad de cobertura, P_c, se le llama *confiabilidad* y a la longitud del intervalo, y por ello a $V_p(\hat{R})$, la *precisión* de la estrategia.

Así como una muestra particular entrega una estimación del porcentaje de favoritismo por un candidato, otra muestra, conformada por otros municipios, otras manzanas u otras personas arroja una estimación diferente. En general, para cada muestra, de la inmensa cantidad teórica de muestras posibles, se tiene una estimación o valor del porcentaje de favoritismo por el candidato. Sobre este marco de todas las estimaciones diferentes, cada una asociada a su muestra, que a su vez tiene una determinada probabilidad [7] de ser extraída, se define confiabilidad como (Särndal et al. 2003) la suma de las probabilidades de las muestras, cuyo intervalo de confianza cubre al valor real.

El Teorema Central de Límite (TCL) afirma que la distribución de los promedios muestrales, tiende hacia una distribución Normal o campana de Gauss con ciertos parámetros, a medida que el tamaño de muestra crece. En tal caso, la probabilidad de cobertura, y con ella la confiabilidad se deja calcular fácilmente, y es igual a $(1-\alpha)$, con α establecido en el valor $z_{1-\frac{\alpha}{2}}$ (de la fórmula (3)) de la tabla de la normal estándar. En el caso de estimación de una razón, no se tiene una afirmación similar a la del TCL para los promedios. La solución propuesta por la teoría estadística es aplicar el TCL a modo de aproximación, con lo que la probabilidad de cobertura, y por ende la confiabilidad es inferior a $(1-\alpha)$. En diseños complejos, por ejemplo de varias etapas y muestras pequeñas la aproximación es tan deficiente, que la verdadera probabilidad de cobertura o confiabilidad es tan baja que hace los resultados inútiles [8] (McManus 2004) (Gawiser & Witt 2002).

Recurriendo de nuevo al símil, meramente teórico, de la inmensa lista de porcentajes estimados, uno por cada muestra posible, se espera que ellos oscilen alrededor del valor real que se pretende estimar, $E_p(\hat{R}) = R$. Es decir, se espera que la estrategia "apunte" a lo que se busca. Si las estimaciones "apuntan a otra parte" se dice que se trata de una estrategia con sesgo. En ese caso P_c la probabilidad de cobertura o confiabilidad será baja. En estrategias no desviadas, es decir sin sesgo, la confiabilidad, dependiendo de la calidad de la aproximación al aplicar el TCL, se acerca a $(1-\alpha)$. Cuando la estrategia tiene sesgo, la confiabilidad decrece en función de la magnitud del sesgo. En muestras grandes, el muestrista debe mantener una estricta vigilancia a fin de no introducir, o en forma más realista,

[7] Conocida, por cumplir la condición de ser muestra probabilística

[8] Una muestra de 1.200 entrevistados en las cuatro principales ciudades del país contaría, en el mejor de los casos, con 50 mujeres de un mismo nivel socio-económico en una ciudad. Con ese minúsculo tamaño de muestra cualquier afirmación sobre preferencia electoral femenina por estrato y ciudad no puede ser confiable.

a fin de controlar la mayor cantidad posible de fuentes de sesgo. El sesgo puede provenir, entre otras fuentes, de errores del marco de muestreo, como por ejemplo la subcobertura[9]. También se produce por errores en el empadronamiento, como por ejemplo el mal tratamiento de las novedades[10]. Originan sesgo, las entrevistas diligenciadas fraudulentamente por el entrevistador, la aplicación de métodos de muestreo sin el debido rigor técnico que ellos exigen[11], la utilización de factores de expansión erróneos o de fórmulas de cálculo equivocadas[12].

Una vez garantizada la mayor confiabilidad posible, el muestrista busca reducir el tamaño del intervalo de confianza a fin de entregar resultados útiles. En el caso particular de una ENFEP no se necesita una muestra estadística para "saber" de antemano, que un determinado candidato obtendrá, por ejemplo, una votación entre el 20 y el 40 por ciento. La tarea del muestrista en una ENFEP es producir intervalos con una longitud inferior a cinco o seis puntos porcentuales. Para el caso de longitud igual a seis y si el porcentaje estimado es, por ejemplo, 34 %, entonces el porcentaje verdadero de favoritismo está, con una alta probabilidad, garantizada por la confiabilidad, entre (34 ± 3) %, es decir entre (31 % y 37 %). Para alcanzar este intervalo de confianza y sobre la base de que se pretende una confiabilidad cercana al 95 %, lo que significa que la constante $z_{1-\frac{\alpha}{2}} = 1{,}96$; que para efectos prácticos se toma igual a 2; se debe proyectar una estrategia que cumpla:

$$\sqrt{V_p(\hat{R})} \leq 0{,}015 \iff V_p(\hat{R}) \leq 0{,}000225$$

Volviendo al símil de la tabla con todas las muestras posibles, cada muestra con su correspondiente estimativo, lo que se pretende es que no haya mucha variación entre las diferentes estimaciones[13]. Para mantener la precisión en los rangos deseados, el muestrista juega, entre otros, con tres aspectos básicos: El diseño de muestra, que es la forma probabilística como selecciona conglomerados y elementos; con la definición del estimador o formas de cálculo y con la definición de los tamaños de muestra[14].

El tamaño de muestra adecuado depende de la configuración del universo de estudio. Cuando un candidato polariza la población en forma tal que casi todos los habitantes de ciertas manzanas lo apoyan, mientras que en otros sectores nadie votaría por él[15], lo conveniente desde el punto de vista de reducir la variabilidad de

[9] Ausencia en mapas o listados de barrios o sectores de la ciudad construidos en los últimos años

[10] Se denominan novedades los casos de múltiples, fuera de universo y no-respuesta.Múltiples: en el mapa aparece una manzana y en la realidad son varias, Fuera de universo: en el mapa aparece una manzana con viviendas y lo que el empadronador encuentra es, por ejemplo, una estación del sistema de transporte masivo, y la No-respuesta cuando, por ejemplo, en un edificio de apartamentos no se obtiene permiso para conocer la cantidad de residentes.

[11] Traslape en conglomerados o estratos, desatención del principio de invarianza, etc.

[12] Estimadores no apropiados

[13] Una estrategia para la ENFEP no sería adecuada, si al estimar el porcentaje de favoritismo muchas muestras arrojan porcentajes del orden del 15 %, muchas otras, de la misma estrategia, señalan favoritismo de alrededor del 35 % y otras tantas entregan tasas de favoritismo alrededor del 60 %.

[14] Se dice *tamaños de muestra*, porque en diseño de dos o más etapas son varios los procesos de selección que se deben realizar.

[15] En tal caso se dice que el candidato genera correlación intraclásica

las estimaciones, es seleccionar muchas manzanas y pocas personas por manzana; método por lo demás costoso frente a la alternativa de conformar la muestra, tomando muchas personas por manzana de algunas pocas manzanas empadronadas.

El objetivo, en el ejemplo numérico que se viene tratando, es entonces establecer un diseño de muestra, unos tamaños muestrales y unos estimadores tales que la varianza del estimador sea menor, por ejemplo a dos diezmilésimos.

En la mayoría de los casos es relativamente complicado establecer límites para la varianza, puesto que se trata de unidades cuadradas. Por ello se acostumbra tratar el tema de la varianza del estimador en forma relativa utilizando el concepto de coeficiente de variación del estimador $CV_p(\hat{R})$, dado, para este trabajo, por:

$$CV_p(\hat{R}) = \frac{\sqrt{V_p(\hat{R})}}{R} \tag{4}$$

lo que en el caso numérico que se viene exponiendo y si la verdadera razón es $R = 0{,}325$ equivale a decir que el $CV_p(\hat{R}) \leq \dfrac{0{,}015}{0{,}325} = 0{,}046 = 4{,}6\,\%$.

En general, se califica la calidad de la precisión, en función del coeficiente de variación, como se muestra en el cuadro 1.

Tabla 1: Calificación de la calidad de la precisión de la estrategia muestral en función del valor del Coeficiente de Variación $CV_p(\hat{R})$

Valor del $CV_p(\hat{R})\,(\,\%)$	Calificación de la precisión
Menor a 2 %	Excelente
Entre 2 % y 4 %	Buena
Entre 4 % y 6 %	Moderada
Entre 6 % y 10 %	Baja
Entre 10 % y 15 %	Para usar sólo con mucho cuidado
Superior a 15 %	No se puede publicar

Para una tasa de favoritismo del 20 % con una estimación de precisión moderada, por ejemplo, $CV = 5{,}2\,\%$, se estaría entonces diciendo que:

$$\sqrt{V_p(\hat{R})} = (R)(CV_p(\hat{R})) = (0{,}2)(0{,}052) = 0{,}0104$$

con lo que el intervalo de confianza tendría a cada lado una longitud igual a $(2)(0{,}0104) = 0{,}0208 = 2{,}1\,\%$. Es decir que cuando se emita un estimativo $\hat{R}$, el verdadero valor estaría con alta probabilidad en el intervalo $[\hat{R} \pm 2{,}1\,\%]$.

2. Construcción de la estrategia muestral

2.1. Varianza de la estrategia $V_p(\hat{R})$

El objetivo es establecer una estrategia muestral que mantenga la varianza de la tasa estimada de favoritismo por debajo de una determinada cota. Sin embargo, para planificar esa estrategia es necesario conocer la tasa de favoritismo, lo cual constituye un círculo vicioso. La solución práctica, aplicada en general y en particular en este ejercicio, es utilizar datos completos de períodos anteriores, como si ellos constituyeran los datos desconocidos del día de hoy. Para realizar estimaciones referentes a las elecciones de 2002 en Colombia se toman los datos de la elección de 1998. En ambas elecciones, estuvo el candidato Horacio Serpa como fuerte competidor por la Presidencia.

El camino que se propone en este trabajo, es el de conformar una base de datos, que combina la información persona a persona del censo nacional de población y vivienda de 1993 (DANE 1996) con información electoral de la primera vuelta de 1998. La información censal contiene la identificación de manzana, sección, sector cartográfico, zona rural o urbana y municipio, y la información electoral permite reproducir los resultados de la cantidad de votantes y la cantidad de personas que, en cada municipio votaron por el candidato Serpa en la primera vuelta de 1998. Para ello se generan aleatoriamente para cada persona las variables y_k, z_k como se señala en (1). Los valores y_k y z_k, así generados, conducen a que la cantidad de votantes y de votos por Serpa son acordes a los resultados reales de 1998, tanto a nivel de municipio[16] como para el total del país, y proveen una base ficticia de distribución de votantes y partidarios de Serpa, por sector, sección y manzana. Esta configuración de datos cumple un importante supuesto pero desatiende otro igualmente importante.

La generación aleatoria de valores y_k, z_k en forma separada e independiente al interior de cada municipio respeta la fuerte correlación intraclásica del conglomerado "Municipio". Sin embargo, la generación aleatoria al interior de los municipios, sin considerar niveles socio-económicos, sexo, edad o niveles culturales de la población, está suponiendo que la votación por Serpa sigue, al interior de los municipios, un patrón de muy baja correlación intraclásica. Es decir, que no se concentra en determinados sectores poblacionales. Para subsanar este defecto, en la parte final del trabajo, se realiza la prueba de la estrategia propuesta, concentrando la votación y el favoritismo en ciertas partes del municipio para producir valores altos de correlación intraclásica entre las secciones cartográficas.

Al utilizar los resultados electorales de 1998 con la base de datos del censo de 1993 se respeta la estructura de manzanas, secciones, sectores, y la clasificación urbano-rural pero no se contempla el crecimiento poblacional de esos cinco años, de forma tal que se reproducen los resultados de votación y favoritismo por Serpa y se supone que las partes no estudiadas, por no disponer de información actualizada, se comportan, sencillamente, como el resto del municipio.

[16]Tan sólo en algún municipio muy pequeño y marginal sucede que la cantidad de adultos en 1993 es menor que la de votantes en 1998.

Con la base de datos así construida se busca la mejor estrategia que cumpla una determinada cota para la varianza de $\hat{R}$. Por tratarse de la estimación de una razón, el cálculo de la varianza de la estimación se obtiene mediante la aproximación de Taylor, y para ello es necesario construir la transformada:

$$u_k = \frac{1}{N_z}(y_k - Rz_k) \tag{5}$$

con lo que, la varianza que se busca queda dada por:

$$V_p(\hat{R}) = \sum \sum_{U_I} \Delta_{Iij} \frac{t_{uU_i}}{\pi_{Ii}} \frac{t_{uU_j}}{\pi_{Ij}} + \sum_{U_I} \frac{V_i}{\pi_{Ii}} \tag{6}$$

donde:

U_I es el conjunto de conglomerados primarios de muestreo (municipios),

$\Delta_{Iij} = \pi_{Iij} - \pi_{Ii}\pi_{Ij}$ con π_{Ii} y π_{Iij} las probabilidades de inclusión de primero y segundo orden del diseño muestral de CPMs,

t_{uU_i} es la suma en el i-ésimo CPM (municipio) de las transformadas, es decir $t_{uU_i} = \sum_{U_i} u_k$

V_i es la varianza al interior del i-ésimo municipio, lo que significa realizar de nuevo el cálculo de la varianza en varias subetapas.

Con los y_k y z_k generados para la población completa se construye la transformada (5), que para el caso individual, asume sólo tres valores:

$$u_k = \begin{cases} 0 & \text{si } z_k = 0 \text{ ya que entonces todo } y_k = 0 \\ \frac{1}{N_z}(1-R) & \text{si } y_k = 1 \text{ y } z_k = 1 \\ \frac{1}{N_z}(-R) & \text{si } y_k = 0 \text{ y } z_k = 1 \end{cases}$$

La suma de los valores u_k al interior del i-ésimo municipio es igual a:

$$\begin{aligned} t_{uU_i} &= \sum_{U_i} u_k = \sum_{U_{yi}} \frac{1}{N_z}(1-R) + \sum_{U_{zi} \cap U_{yi}^c} \frac{1}{N_z}(-R) \\ &= \frac{N_{zi}}{N_z}(R_i - R) \end{aligned} \tag{7}$$

donde N_{zi} es la cantidad de votos emitidos en el municipio, t_z la cantidad nacional de votos, R_i la proporción de favoritismo por Serpa en el municipio y R la tasa nacional de favoritismo por el mismo candidato. Este total se hace igual a cero, si la tasa municipal de favoritismo R_i es igual a la tasa nacional R, lo que ocasiona que algunos municipios grandes aporten poco a la varianza total de la estrategia, mientras que otros, con menos votación, pero con una marcada tendencia a favor o en contra de Serpa, logran valores, positivos o negativos, lejanos de cero.

2.2. Primer escenario: muestreo aleatorio simple de municipios

Como ya se mencionó, no hay posibilidad, por carencia del necesario marco de muestreo, de realizar un muestreo directo de elementos. Pero, como es sabido, la varianza de la estrategia crece a medida que se adicionan etapas al diseño. La opción es intentar un diseño con tan pocas etapas como sea viable. Se comienza por definir el conglomerado muestral de primer orden, que conviene estudiar, y puesto que, al interior del conglomerado es necesario realizar un empadronamiento se busca, en consecuencia un conglomerado de tamaño pequeño. Para la definición del conglomerado primario de muestreo, el menor nivel, sobre el que se tiene información idónea es el municipio, que es el CPM escogido en esta propuesta. La primera idea de diseño muestral es, realizar una muestra aleatoria simple de municipios. La fórmula de la varianza debida a la primera etapa, que le corresponde a este diseño es:

$$V_{ET1-MAS}(\hat{R}) = \frac{N_I^2}{n_I}(1 - \frac{n_I}{N_I})S^2_{t_u U_I}$$

$$= \frac{N_I^2}{n_I}(1 - \frac{n_I}{N_I})\frac{1}{N_I - 1}\sum_{U_I}(t_{uU_i} - \bar{t}_{U_I})^2$$

pero como

$$\bar{t}_{U_I} = \frac{\sum_U u_k}{N_I} = 0 \quad \Rightarrow \quad S^2_{t_u U_I} = \frac{1}{N_I - 1}\sum_{U_I}(t_{uU_i})^2$$

con lo que los municipios que más aportan a la varianza de la estrategia son aquellos con mayor valor absoluto de t_{uU_I}. Con este diseño se requeriría una muestra de cerca de 600 municipios para alcanzar un CV cercano al 4 % (Véase cuadro 2).

Tabla 2: Coeficiente de variación $CV_p(\hat{R})$ alcanzado por la primera etapa según el tamaño de muestra propuesto utilizando un diseño MAS en la primera etapa

Tamaño de la muestra de la primera etapa Cantidad de municipios a seleccionar	$CV_{1-MAS}(\hat{R})$ (%)
720	3,02
585	4,04
475	5,03
385	6,03
315	7,03
260	8,03

La dispersión de los valores $|t_{uU_I}|$ es tan alta que los resultados conducen rápidamente a la necesidad de considerar estrategias diferentes a la del MAS para

la primera etapa. Para el diseño muestral de la primera etapa, es decir, para la selección de municipios, se tienen entonces dos posibilidades: realizar una muestra con probabilidad proporcional al tamaño de $\left|t_{uU_I}\right|$ o estratificar los municipios. El diseño P.P.T. es tenido en cuenta y resulta, desde el punto de vista de la varianza, ligeramente mejor que la estrategia de crear estratos, pero en la práctica presenta complicaciones operativas que no se tienen cuando se escoge la opción del diseño estratificado.

2.3. Segundo escenario: estratificación de municipios

La mayor fuente de variación, para el caso de la estimación de la razón con diseño multietápico, se origina en la fuerte asimetría de los totales $\left|t_{uU_I}\right|$, (ver (7)), de los algo más de mil municipios del país. Hay valores muy grandes de $\left|t_{uU_I}\right|$, que superan las 500 millonésimas hasta Bogotá, en la que $t_{uU_I} = 11{,}684$ millonésimas. Para ese grupo se obtendría una importante reducción de la varianza del estimador, si se reúnen en un estrato, en el que se estudian todos los municipios que lo conforman. Los valores altos de $\left|t_{uU_I}\right|$ corresponden a municipios en los que se combinan dos aspectos: un tamaño amplío y un comportamiento de favoritismo por el candidato Serpa diferente al porcentaje nacional. Nótese que un municipio, por grande que sea, si se comporta porcentualmente como el total del país, es decir $R_i = R$, no aporta a la varianza del estimador, puesto que su suma t_{uU_i} se vuelve cero, y no hace parte de este primer estrato.

Hay valores de $\left|t_{uU_I}\right|$ más modestos que los mencionados anteriormente, que oscilan entre dos y 500 millonésimas, que podrían dar origen a uno o más estratos de municipios. Por último hay muchos valores de $\left|t_{uU_I}\right|$ muy cercanos a cero, desde dos millonésimas hasta fracciones de millonésimas, que aportan muy poco a la varianza general. De este grupo de municipios se puede seleccionar sólo unos muy pocos para reducir costos, sin incrementar en gran medida la varianza del estimador. La propuesta metodológica es, en conclusión, aplicar un diseño estratificado del tipo IF - ESTMAS - UNO, es decir, se investigan todos los municipios del primer estrato, se extraen muestras MAS en los estratos intermedios y en el último estrato se extrae un único municipio.

Para estratificar se trabaja primero con una variación al método propuesto por Hidiroglou (Hidiroglou 1986) para la conformación de un estrato de inclusión forzosa y otro de diseño MAS. La ganancia de precisión, respecto al escenario MAS, es ya muy importante. Para conseguir un CV de primera etapa del 5 %, el método sugerido como variación al propuesto por Hidiroglou pide un tamaño de muestra de 55 municipios, mientras que para alcanzar esa precisión, el diseño MAS exige $n = 480$.

Se prueba luego, en forma análoga, una variación al método de Lavallée (Lavallée & Hidiroglou 1988) para la conformación de un estrato de inclusión forzosa y varios de diseño MAS. Sin embargo esta variación no contempla la posibilidad de un último estrato con un único elemento en la muestra. Se procede entonces a la aplicación de un método de iteración computacional de cálculo de varianza de primera etapa, variando las configuraciones de estratificación. El mecanismo para

determinar la configuración de estratificación que provee la menor varianza del estimador de la tasa de favoritismo, es el siguiente:

1. Se ordenan los registros de los 1016 municipios en forma descendente respecto al cuadrado de la suma de sus valores de la transformada u_k. Es decir se ordenan los municipios en forma descendente respecto a:

$$t_{uU_i}^2 = \sum_{U_i} \left(\frac{1}{t_z}(y_k - Rz_k) \right)^2$$

2. Para un tamaño global de muestra n_I, se calcula la varianza, debida a la primera etapa, que genera la estratificación construida de la siguiente manera:

   - Un primer estrato con diseño de inclusión forzosa de tamaño N_{IF}
   - Un segundo estrato con diseño MAS(N_{I2}, n_{I2}), y
   - Un tercer estrato con diseño MAS($N_{I3}, 1$)

   En este primer ejercicio, con tres estratos, la varianza del estimador de la razón depende de tres parámetros: El tamaño de muestra n_I, el tamaño del estrato de inclusión forzosa N_{IF} con lo que, por diferencia, queda definido el tamaño $n_{I2} = n_I - N_{IF} - 1$, y el tamaño del segundo estrato N_{I2} que determina el tamaño $N_{I3} = 1016 - N_{IF} - N_{I2}$.

3. Una vez realizados los cálculos de varianza para combinaciones de los tres parámetros se escoge aquella configuración que para un tamaño de muestra produce la menor varianza.

El largo trabajo computacional se recompensa con la fuerte reducción alcanzada para la varianza del estimador. La varianza se reduce a la cuarta parte respecto al caso MAS, como se puede observar en la tabla 3.

El siguiente paso es considerar la configuración en cuatro estratos y compararla con la de tres estratos[17]. En tal caso se tienen más parámetros y por ende más cálculos que realizar, pero dentro de la misma lógica de programación. Luego se estudia el caso de cinco estratos. El crecimiento de la cantidad de parámetros hace que la cantidad de cálculos crezca en forma exponencial, pero sigue siempre idéntica estrategia de programación.

El resultado es que con tres estratos se mejora bastante la propuesta basada en la variación al método de Hidiroglou, con cuatro estratos se obtiene una leve ganancia frente a la configuración con tres estratos, y con cinco estratos crece el grado de complejidad, mientras la ganancia, en términos de varianza es muy pequeña. La decisión final de esta propuesta es adoptar el plan de cuatro estratos.

El ejercicio arroja una varianza debida a la primera etapa, equivalente a un $CV_p(\hat{R}) = 3{,}8\,\%$, tomando una muestra de ochenta municipios, distribuida así:

[17] El cálculo de las varianzas variando configuraciones y tamaños de muestra en cuatro estratos tarda algo más de dos horas, realizando cálculos con el paquete de procesamiento estadístico SAS versión 8.2- Computador Pentium 4R- CPU 2,6 GHz, 512 MB RAM.

Tabla 3: Coeficiente de variación $CV_p(\hat{R})$ alcanzado por la primera etapa, según el tamaño de muestra propuesto, cuando se utilizan diseños MAS, y ESTMAS con tres, cuatro y cinco estratos en la primera etapa

n_I	MAS	E=3	E=4	E=5
50	20,7	5,63	5,39	5,31
60	18,8	4,89	4,72	4,67
70	17,3	4,31	4,22	4,21
80	16,1	3,92	3,85	3,83
90	15,1	3,67	3,51	3,53
100	14,3	3,35	3,27	3,29
110	13,5	3,14	3,06	3,04
120	12,9	2,96	2,86	2,86
130	12,3	2,81	2,70	2,66
140	11,8	2,64	2,51	2,52
150	11,3	2,50	2,39	2,36

- Un primer estrato de inclusión forzosa con 21 municipios, que contempla el 45 % de los votantes del país.
- Un segundo estrato con 144 municipios, de los cuales se estudian 44 (uno de cada tres) y que recogen el 22 % de la votación nacional.
- Un tercer estrato con 610 municipios, de los cuales se visitan catorce (aprox. dos de cada cien) y que aportan el 25 % de los votos.
- El último estrato con 241 municipios, que totalizan el 8 % de la votación, y de ellos sólo uno será seleccionado para la muestra de la primera etapa.

El estrato de inclusión forzosa recoge los principales municipios del país, aunque al final de la lista aparecen algunas sorpresas y faltan otros, que si el criterio fuera sólo tamaño, allí deberían aparecer, pero como se señaló anteriormente, presentan un porcentaje similar al nacional, que los convierte en poco interesantes desde la perspectiva muestral. La lista de los municipios que conforman el estrato es: Bogotá, Cali, Buenaventura, Medellín, Envigado, Bello, Itagüí, Barranquilla, Soledad, Bucaramanga, Barrancabermeja, Pereira, Dosquebradas, Manizales, Cartagena, Montería, Valledupar, Sincelejo, Villavicencio, Quibdó y Pasto.

2.4. Muestra al interior de los municipios

En los municipios no es viable la construcción de un marco de personas mayores de 18 años, lo que obliga a pensar en diseños en varias etapas y con tan pocas etapas, como sea posible. Sin embargo, se debe considerar, a la vez, otro aspecto fundamental, el costo. Se construyen fácilmente ejemplos en los que se obtiene igual varianza, cuando se estudian muchas manzanas y pocas personas por manzana, que

cuando se toman muestras con pocas manzanas y muchas personas por manzana, sin embargo el costo de las dos estrategias puede ser muy diferente. El costo global de una muestra en varias etapas depende de dos costos bien diferentes, el costo de construcción del marco para la última etapa y el costo de entrevista para la medición propiamente dicha. Para este ejercicio se aplica un costo C_1 para la construcción de la lista de una manzana de tamaño promedio y para la realización de diez entrevistas directas y efectivas de preferencia electoral[18].

En la mayoría de las ciudades del primer estrato no parece conveniente pasar directamente a la selección de manzanas, por los costos asociados a los desplazamientos entre ellas. Aunque la inclusión de una etapa adicional genera mayor varianza, para reducir dispersión en cada una de esas ciudades, se propone seleccionar primero sectores cartográficos, mediante el algoritmo de Fan-Muller-Rezucha. Dentro de los sectores seleccionados escoger manzanas, con el mismo algoritmo. Construir el padrón en cada manzana de la muestra, para escoger de allí, también con el mismo algoritmo, la muestra de personas a entrevistar. En los municipios de los restantes tres estratos la propuesta es seleccionar directamente manzanas y en la siguiente etapa seleccionar personas. Se llega de esta manera a la propuesta de una estrategia muestral estratificada, con un estrato de inclusión forzosa y diseño, a su interior en tres etapas. Otros tres estratos con diseño en tres etapas, selección de municipios, mediante MAS, selección de manzanas, mediante MAS y selección de personas, también con MAS, es decir, diseño MAS^3 (Bautista 1998).

2.5. Resultado final: Diseño muestral para la elección de 2002

El resultado de este ejercicio establece que el diseño final de muestra en cuatro estratos y tres etapas para la estimación de la tasa de favoritismo electoral en Colombia en la elección de 2002 para alcanzar una precisión equivalente a $CV_p(\hat{R}) = 5{,}1\,\%$ queda conformada así:

- **Grandes ciudades:** 21 de 21 municipios; uno de cada 20 sectores cartográficos, mínimo dos por municipio, 60 % de las manzanas por sector y una de cada 25 personas por manzana.
- **Ciudades intermedias:** 44 de 144 municipios, 10 % de las manzanas por municipio y una de cada 25 personas por manzana.
- **Municipios pequeños:** 14 de 610 municipios, 60 % de las manzanas por municipio y una de cada 25 personas por manzana.
- **Municipios muy pequeños y alejados:** 1 de 241 municipios, 60 % de las manzanas por municipio y una de cada 25 personas por manzana.
- **Total Nacional:** 80 municipios, 106 de los 2134 sectores de las 21 ciudades, aproximadamente 6.200 manzanas y alrededor de 15.000 personas.

[18] La equivalencia una manzana empadronada cuesta lo mismo que diez entrevistas efectivas, es un parámetro determinante de los resultados finales obtenidos.

De esta muestra se puede señalar:

El tercer estrato es el que más aporta a la varianza global de la estimación. Un aumento de la cantidad de municipios a seleccionar puede elevar mucho los costos operativos. Sin embargo, se podría intentar obtener alguna ventaja mediante la construcción de "rutas", es decir, la reunión dentro de un mismo conglomerado de municipios pequeños con cercanía geográfica. Esto implicaría una etapa adicional, en ese estrato, y habría que evaluar con cuidado, si la ganancia global de precisión tiene relación con el incremento de costos.

Los ensayos realizados en el sentido de incrementar la muestra de municipios del último estrato muestran que con más de un municipio no se aporta prácticamente nada al mejoramiento de la varianza global de la estimación.

La decisión de tomar una proporción tan pequeña de sectores en las grandes ciudades, uno de cada veinte, parece inadecuada si en la elección presidencial se presenta una fuerte concentración de opinión por sectores. De hecho, muchos sectores son homogéneos en el sentido de que su población es socio-económicamente del mismo nivel, toda ella es pobre, media o de nivel alto. Si algún candidato presidencial logra agrupar favoritismo y rechazo en forma marcada según el nivel socio-económico, la muestra de sectores debería ser un poco mayor.

La proporción de manzanas por sector y municipio parece alta, a la vez que la proporción de personas por manzana es relativamente baja, lo que implica que el costo de construcción del padrón está siendo desaprovechado por la cantidad baja de entrevistas por manzana. La razón de esta decisión se basa en el supuesto de que la correlación intraclásica, es decir la homogeneidad de opinión al interior de la manzana puede ser alta, sin embargo mucho más determinante y cierto es el supuesto de alta correlación intraclásica en lo referente a si se participa o no en el comicio electoral.

Aunque se trataría de una variación al diseño, se puede pensar que con un único padrón y siguiendo un plan de muestras replicadas en varias fases y traslapando algunas partes de las muestras, se podrían hacer mediciones de opinión electoral en seis a ocho momentos diferentes a modo de seguimiento en panel; en tal caso habría necesidad de ampliar un poco la muestra de manzanas para compensar el efecto que sobre la varianza tiene el hecho de aplicar un diseño en fases.

2.6. Prueba de la estrategia propuesta

Para probar la calidad de la muestra así diseñada, se procede al siguiente ejercicio: Se utiliza la información municipal de votación y de favoritismo por Álvaro Uribe en 2002 para generar una base de datos similar a la Serpa 1998 generando aleatoriamente para cada persona si votó o no y si lo hizo o no por Uribe en 2002. De esa base se retiran las personas pertenecientes a sectores rurales, previa construcción del respectivo factor de ajuste[19]. Sobre ese universo así establecido, se aplica la estratificación, los tamaños y las formas de selección establecidas en

[19]El supuesto que sustenta esta decisión es que el comportamiento rural de cada municipio es similar al urbano de ese mismo municipio.

la propuesta planteada.

Se procede entonces a realizar en forma computacional (véase anexo 2), quinientas repeticiones independientes del proceso completo, que abarca desde la selección de municipios, la selección de personas [20] y la estimación del porcentaje de votos que según la muestra le corresponden a los candidatos. r Los resultados obtenidos, sabiendo que la tasa final de favoritismo con la que ganó A. Uribe en 2002 en el país fue 53,87 %, son los siguientes:

- Cantidad de repeticiones independientes = 500
- Promedio de las estimaciones de las 500 réplicas = 0,5383= 53,83 %
- Porcentaje de réplicas con estimación superior al 50 % = 96 %
- Varianza estimada de la estrategia = 0.000443
- Confiabilidad estimada, es decir porcentaje de réplicas en las que

$$\begin{aligned} 0,5387 = R_y \in & \left[\hat{R}_y \pm z_{1-\frac{\alpha}{2}}\sqrt{\hat{V}_p(\hat{R}_y)}\right] \\ \in & \left[\hat{R}_y \pm (2)\sqrt{0{,}000443}\right] \\ \in & \left[\hat{R}_y \pm 0,042084\right] \qquad = 94,8\,\% \end{aligned}$$

- Coeficiente de variación estimado c.v.e = 3,9 %
- Cantidad media de manzanas a enlistar = 6.110
- Cantidad media de personas a entrevistar = 14.530

Es evidente que sólo, con muy mala suerte se obtendría una muestra con la que se afirmaría, antes de las elecciones, que habría segunda vuelta. En el 96 % de los casos, la muestra así diseñada habría pronosticado el triunfo inmediato del candidato Uribe. La confiabilidad estimada es casi del 95 %, lo que no es necesariamente sorprendente, puesto que, se trata de un ejercicio de cómputo, en el que no se introducen los sesgos frecuentes en los operativos de campo. La precisión obtenida es equivalente a un c.v.e. de 3,9 % que dista algo del CV programado y equivalente a 5,1 %. La causa de esta diferencia radica en que la muestra diseñada utiliza como base la información referente al candidato H. Serpa, quien en la primera vuelta de 1998 obtuvo sólo 34,3 % del favoritismo, mientras que en el ejercicio presentado, el candidato Uribe obtuvo el 53,8 % del favoritismo. Puesto que tanto el CV como su estimación el c.v.e son medidas relativas, es decir, tienen como denominador la tasa de favoritismo, ellas toman valores bajos para tasas altas y valores altos

[20] El proceso de seleccionar aleatoriamente 59 municipios de los estratos 2, 3 y 4; seleccionar 106 secciones, cerca de seis mil manzanas y alrededor de 15.000 personas, siempre con el algoritmo de Fan-Muller-Rezucha, y realizar la estimación pedida se realiza en 51.8 segundos. Las 500 repeticiones de este proceso tarda 7,2 horas, con las especificaciones de hardware y software señaladas anteriormente.

para tasas pequeñas[21]. La muestra propuesta resulta insuficiente para estimar con confiabilidad y precisión la tasa de favoritismo de los candidatos que ocuparon el tercer y cuarto lugar en la elección de 2002. Es natural, que tratándose de porcentajes tan bajos, 6,3 % y 6,0 % respectivamente, las muestras necesarias sean considerablemente grandes[22].

La muestra propuesta tiene el inconveniente, más teórico que práctico, de no entregar un tamaño de muestra relativamente constante, que haga posible una aproximación al costo total del operativo. Puesto que el plan muestral toma porcentajes de manzanas en los municipios seleccionados y ellos varían en cada muestra de primera etapa, la cantidad de manzanas a empadronar termina siendo variable. De igual manera, la cantidad de personas a entrevistar depende del tamaño de las manzanas, que aleatoriamente se seleccionen en la muestra de la segunda etapa. El ejercicio realizado señala que en el 76 % de los casos la cantidad de manzanas a empadronar es una cantidad entre 5.500 y 6.700 y en el 80 % de las réplicas se deben entrevistar entre trece y dieciséis mil personas.

2.7. Aplicación de la metodología propuesta para las elecciones presidenciales de 2006

Si en 2006, el Presidente Álvaro Uribe, vuelve a ser candidato a la Presidencia de la República, se estaría en un caso similar a lo sucedido con Serpa 1998-2002. Así como se utilizan los datos de Serpa 1998, para el diseño de la muestra 2002, se pueden utilizar los datos de Uribe 2002, para el diseño de una posible muestra para una ENFEP-2006. Se siguen entonces los mismos pasos y se llega al siguiente resultado global, el que para poder ser considerado como plan muestral final, debería ser trabajado y presentado con mayor detalle.

Se particiona el conjunto de municipios del país en cuatro estratos, el primero con diseño de inclusión forzosa, y tres de inclusión probabilística. Las elecciones de 2002 estuvieron marcadas, a diferencia de lo sucedido en los comicios anteriores, por una fuerte polarización del favoritismo en los municipios. Esa polarización genera un fuerte crecimiento del estrato de inclusión forzosa, una importante reducción del segundo estrato y una mayor concentración muestral en él. Es decir, en la muestra de la ENFEP-2002 eran necesarios, en el segundo estrato, 44 de 144 municipios, algo más de uno por cada tres, para la ENFEP-2006 se necesitan 19 de 42 municipios. El estrato de inclusión forzosa que antes estaba conformado por 21 municipios, contiene ahora 38 municipios, lo que significa un crecimiento del 80 %. Para las elecciones el 2006 entrarían en el diseño muestral que aquí se propone, de manera segura en la muestra, los municipios: Bogotá, Cali, Buenaventura, Tulúa, Cartago, Medellín Envigado, Bello, Itagüí, Rionegro, Barranquilla, Soledad, Cartagena, Cúcuta, Bucaramanga, Girón, Floridablanca, Barrancabermeja, Manizales, Pereira, Dosquebradas, Santa Rosa de Cabal, Armenia, Santa Marta,

[21] Realizado el mismo ejercicio para estimar los resultados del candidato Serpa se obtuvo un promedio de 31,5 % contra 32,4 % realmente obtenido y un c.v.e de 5,9 %.

[22] Aun mayor deben ser las muestras necesarias para la estimación de la tasa nacional de favoritismo de candidatos al Senado de la República

Ciénaga, Riohacha, Maicao, Montería, Sahagún, Valledupar, Sincelejo, Quibdó, Ibagué, Soacha, Villavicencio, Sogamoso, Puerto Tejada y Pasto.

El tercer estrato contiene ahora 594 municipios, y de él se extraen 27 municipios. En este estrato se encuentran aún algunas capitales departamentales importantes, como Neiva, Popayán y Florencia. El último estrato contiene los 342 municipios más pequeños y de él se extrae un único municipio. En total, la muestra para la primera etapa de la ENFEP-2006 es de 85 municipios.

El crecimiento de la muestra en los dos primeros estratos implica un crecimiento en la cantidad de manzanas a empadronar. Los municipios que componen el primer estrato tienen tamaños muy diferentes lo que sugiere un tratamiento particular de la cantidad de sectores a seleccionar en cada municipio. Se crean entonces cinco grupos de municipios. Bogotá, que conforma el primer grupo y ciudad, para la que se propone una muestra de tres por cada veinte sectores cartográficos. Cali, que conforma el segundo grupo, para la que, en la muestra se toma el 20 % de los sectores. Luego los municipios con más de 70 sectores, en ellos la muestra es el 25 % de sus sectores. El grupo cuatro lo conforman los municipios que tienen entre quince y setenta sectores. En ellos la muestra es la mitad de sus sectores. El quinto grupo, aquellos municipios con menos de quince sectores, en los que todos sus sectores hacen parte de la muestra de la segunda etapa.

Para los sectores de los municipios del primer estrato se propone, entonces, una muestra en la tercera etapa, equivalente a dos de cada veinticinco manzanas. En los municipios seleccionados en el estrato dos se toma una muestra de manzanas, de tamaño equivalente a empadronar tres de cada veinte. En los municipios seleccionados de los estratos tres y cuatro, se empadronan siempre la mitad de las manzanas residenciales. Para todos los casos, la propuesta global, que bien podría ser afinada a fin de reducir costos, es tomar una de cada treinta personas, o lo que equivale a un promedio de 2,5 personas por manzana. Con estos valores de tamaños de muestra se concluye en una muestra global de cerca de 6.400 manzanas y 15.800 personas a entrevistar. Con esta propuesta se consigue un CV de 2,8 % para el porcentaje de 53,9 % que obtuvo el candidato Uribe en 2002. Se trata, sin duda, de tamaños conservadores, y el coeficiente de variación propuesto puede ser calificado de ambicioso. Sin embargo, vale la pena considerar con anticipación algunos comportamientos políticos, que tienen efecto estadístico importante, y que pueden terminar señalando dichos tamaños de muestra como apropiados.

Es posible que el candidato Uribe no obtenga, en la primera vuelta de la elección de 2006, una votación tan voluminosa como en 2002. Si el favoritismo llega, en esta elección alrededor del 40 %, se estaría ante un coeficiente de variación cercano al 4 %. De otra parte, se debe considerar que si la polarización política de los municipios es atribuible, en buena parte al candidato Uribe y sus propuestas políticas, dicha polarización se puede presentar también y en forma marcada entre diferentes niveles socio-económicos. Este fenómeno puede ser mucho más fuerte, dependiendo del o los candidatos más importantes que se opongan a él en la elección. Desde el punto de vista estadístico, el efecto de dicha polarización es la elevación de la correlación intraclásica a nivel de sectores cartográficos y de manzanas. En consecuencia es necesario tomar muestras con más manzanas y pocas personas por

manzana, como la propuesta que aquí se discute. Obviamente, si se anticipa que ninguno de estos dos fenómenos se presentará en la elección, podrían hacerse alguna reducciones importantes en cantidad de manzanas a empadronar y personas a entrevistar.

A. Anexo 1

Tabla 4: Cantidad de municipios según porcentaje de votos para Andrés Pastrana en la segunda vuelta de 1994 (filas) cruzado con sus resultados en la 2a vuelta 1998 (columnas)

	Total	68,4 % ó más	54,7 % a 68,3 %	37,1 % a 54,6 %	0 a 37 %
Total	1019	250	250	250	269
71,7 % ó más	250	221	28	1	
48,5 % a 71,6 %	250	28	172	50	
31,5 % a 48,4 %	250	1	47	150	52
0 a 31,4 %	269		3	49	217

Tabla 5: Cantidad de municipios según porcentaje de votos para Horacio Serpa en la segunda vuelta de 1998 (filas) cruzado con sus resultados en 2002 (columnas)

	Total	0 a 18,9 %	19 % a 35,2 %	35,3 % a 54,9 %	55 % ó más
Total	1019	250	250	250	269
0 a 25,1 %	250	189	53	8	
25,2 % a 45,2 %	250	47	125	71	7
45,3 % a 62,9 %	250	9	50	106	85
63 % ó más	269	5	22	65	177

Tabla 6: Cantidad de municipios según porcentaje de votos para Andrés Pastrana en la segunda vuelta de 1994 (filas) cruzado con los resultados de Álvaro Uribe en 2002 (columnas)

	Total	62,2 % ó más	46,2 % a 62.1 %	32,1 % a 46,1 %	0 a 32 %
Total	1019	250	250	250	269
71,7 % ó más	250	154	63	27	6
48,5 % a 71,6 %	250	59	98	62	31
31,5 % a 48,4 %	250	24	55	91	80
0 a 31,4 %	269	13	34	70	152

B. Anexo 2

Lógica de programación para la generación de quinientas repeticiones de selección de muestra y estimación de la tasa de favoritismo, para la elección presidencial de 2002

Paso 1. Se fija que en los 120 municipios más grandes, se presenta el fenómeno de correlación intraclásica en las secciones cartográficas. El 30 % de las secciones de esos 120 municipios se denominan de tipo **a** y el resto, de tipo **b**. En en los demás municipios todas las secciones son de tipo **c**.

Paso 2. Para cada uno de los 19.109.852 registros se genera aleatoriamente un valor z_k, igual cero o uno de la siguiente forma: si el individuo pertenece a una sección tipo **a**, se hace $z_k = 1$ con probabilidad igual al cociente entre el 23 % de la votación total del municipio en 2002 y la población mayor de 18 años en el municipio.Si el registro pertenece a una sección tipo **b**, se hace $z_k = 1$ con probabilidad igual al cociente entre el 77 % de la votación total y la población del municipio. Si el individuo pertenece a una sección tipo **c**, se hace $z_k = 1$ con probabilidad igual al cociente entre votación y población total del municipio.

Paso 3. Para cada uno de los registros se genera aleatoriamente un valor y_k, igual cero o uno concentrando el 15 % de la votación por Uribe en las secciones tipo **a** y el 85 % en las secciones tipo **b**. Si el registro es de una sección tipo **c**, se hace $y_k = 1$ con probabilidad igual al cociente entre la votación por Uribe en 2002 y la cantidad de votos válidos en ese municipio en dicha elección.

Paso 4. Para cada municipio se establecen los valores de los tamaños muestrales sectxmpio, manzxsect, manzxmpio y persxmanz, de acuerdo al plan muestral propuesto, se crea el factor de corrección por ruralidad y se eliminan los datos correspondientes a las zonas rurales.

Paso 5. Se establece para cada municipio, cada sector y cada manzana el tamaño específico de muestra que le correspondería si fuera seleccionado, ordena los registros siguiendo la jerarquía de selección: estrato, municipio, sector, manzana y persona; y procede a la numeración, necesaria para poder aplicar el algoritmo de Fan-Muller-Rezucha (Särndal et al. 2003), al interior de cada una de las cinco jerarquías[23].

Paso 6. Se elabora una rutina macro de selección Fan-Muller-Rezucha para MAS^3 denominada sel_mas_3, que efectúa:

- Para los municipios del primer estrato realiza la selección aleatoria de sectores cartográficos.
- Para los estratos dos, tres y cuatro realiza la selección de municipios.

[23] Con las especificaciones de software y hardware dadas anteriormente, el proceso que contempla estos primeros cinco pasos preparatorios dura 4,98 horas.

- Para los sectores y municipios seleccionados realiza la selección de manzanas.
- Para las manzanas seleccionadas realiza la selección de personas.
- Para la muestra seleccionada calcula la tasa de favoritismo, utilizando como factor de expansión el producto del factor de corrección por ruralidad por el factor teórico correspondiente al diseño $EST-MAS^3$.

$$f_{ke} = fcr_{ie}\,\frac{N_{Ie}}{n_{Ie}}\,\frac{N_{ie}}{n_{ie}}\,\frac{N_{iqe}}{n_{iqe}}$$

Paso 7. Se elabora una rutina macro, de nombre simula_K, que crea una base de resultados, para un parámetro K dado, invoca K-veces a la macro sel_mas_3 y adiciona la tasa estimada a la base de resultados.

Paso 8. Se invoca la macro simula_K, con K = 500.

Bibliografía

Bautista, L. (1998), *Diseños de muestreo estadístico*, Universidad Nacional de Colombia, Bogotá.

Bautista, L. (2000), Diseño y desarrollo de encuestas, *in* 'Simposio Colombiano de Estadística', Universidad Nacional de Colombia, San Andrés.

Bautista, L. & Pacheco, P. (1989), 'Análisis de la evolución del comportamiento electoral departamental en los últimos años. una aplicación de los métodos factoriales al estudio de series temporales cortas', *Revista Colombiana de Estadística* **19**(2), 94–112.

Biemer, P., Folsom, R., Kulka, R., Lesler, J., Shah, B. & Weeks, M. (2003), 'An evaluation of procedures and operations used by the voter news service for the 2000 presidential election public', *Public Opinion* **67**(Q3), 32–44.

DANE (1996), *XVI Censo nacional de población y V de vivienda*, DANE, Bogotá.

Gawiser, S. R. & Witt, E. (2002), '20 questions a journalist should ask about poll results', *National Council on Public Polls* .

Hidiroglou, M. A. (1986), 'The construction of a self-representing stratum of large units in survey design', *The American Statistician* **40**, 27–31.

Lavallée, P. & Hidiroglou, M. (1988), 'On the stratification of skewed populations', *Survey Methodology* **14**, 33–43.

McManus, J. (2004), 'How reliable are political polls?'.
*http://www.stanford.edu/group/gradethenews

RNEC (1994), *Elecciones presidenciales de 1994 en Colombia*, Registraduría Nacional del Estado Civil, Bogotá.

RNEC (1998), *Elecciones presidenciales de 1998 en Colombia*, Registraduría Nacional del Estado Civil, Bogotá.

RNEC (2002), *Elecciones presidenciales de 2002 en Colombia*, Registraduría Nacional del Estado Civil, Bogotá.

Särndal, C. E., Swensson, B. & Wretman, J. (2003), *Model Assisted Survey Sampling*, 2 edn, Springer Verlag, New York.