Informe 2da parte de las Prácticas Curriculares en el Cerro

Calibración de ponderadores de unidades muestreadas

Juan Manuel Saibene & Joaquín Viola

FCEA FACULTAD DE CIENCIAS ECONÓMICAS Y DE ADMINISTRACIÓN

UNIVERSIDAD DE LA REPÚBLICA URUGUAY

Informe 2da parte de las Prácticas Curriculares en el Cerro

Juan Manuel Saibene & Joaquín Viola

Noviembre de 2023

Índice

1. Introducción

En el primer semestre del año 2023, se hizo un trabajo de prácticas curriculares orientada a estudiantes de la Licenciatura en Estadística de la Facultad de Ciencias Económicas y de Administración, en el marco del proyecto "Circuito Limpio" de la cooperativa de trabajo "La Paloma" y el programa APEX del Cerro de la Universidad de la República. La Cooperativa La Paloma es una cooperativa de trabajo de clasificadores de residuos ubicada en el barrio del Cerro, tiene 9 años y es producto de la ley de envases. Al día de hoy uno de los desafíos que tiene es la cantidad de material para trabajar que llega en mal estado, en parte, por la mala clasificación de los materiales reciclables por parte de los hogares. En busca de contar con mejor material de trabajo, es decir, que lleguen más residuos a su planta y en mejores condiciones es que se embarcan en el proyecto "Circuito Limpio", también como una forma de involucrar a los residentes del barrio en el proceso de reciclaje. El programa Apex de la UdelaR es uno de los programas que tiene la Universidade de la República para desarrollar la función de **Extensión** de la misma, esta función es la encargada de generar el relacionamiento e intercambio entre la UdelaR y la sociedad.

La primer parte de la práctica currricular consistó de la realización de una encuesta sobre hábitos del manejo de residuos y prácticas de reciclaje en los hogares del barrio "El Tobogán" y una parte de la "Villa del Cerro". El mecanismo para el muestreo fue diferente en cada uno de los barrios mencionados. En "El Tobogán" hay 10 zonas censales, por lo que se hizo un muestreo estratificado, visitando 6 casas de cada una de las zonas del barrio, estas 6 casas fueron seleccionadas a través de un diseño sistemático.

En "La Villa del Cerro" tenemos 76 zonas, por lo que se hizo un diseño en dos etapas, en la primera se seleccionaron las 30 zonas a las que debíamos ir, a través de un diseño con probabilidad proporsional al tamaño, según la cantidad de hogares existentes en la zona, y en la segunda etapa se hizo un diseño sistemático, yendo a 6 hogares en cada una de las zonas seleccionadas en la primer etapa.

En esta segunda parte del trabajo, empezando por los hogares de "La Villa del Cerro" buscaremos en primer lugar calcular los ponderadores originales de los hogares, luego ajustarlos por no respuesta, y por último calibrarlos por alguna variables conocidas para los hogares encuestados y para el total de la población (usabdo como marco el censo del año 2011). Estas variables serán el sexo y grupo etario de los integrantes del hogar. Luego repetir el procedimiento para los hogares de "El Tobogán" si hay datos disponibles de confianza para los totales poblaciones.

2. Obtención de la Muestra y No Respuesta

En esta sección se pretende detallar los métodos con los que se obtuvo la muestra de hogares, mencionados en la introducción, y los métodos que se utilizarán para ajustar los ponderadores por no respuesta y luego para calibrar los mismos.

2.1. Obtención de la muestra y ponderadores

Como se mencionó antes, en "Villa del Cerro" la muestra fue obtenida mediante un diseño en dos etapas. Los diseños en varias etapas se utilizan sobre poblaciones agrupadas por estratos o conglomerados. En este caso tenemos a los hogares agrupados en manzanas, que coinciden con la última clasificación por la que están identificadas las viviendas en el Instituto Nacional de Estadística.

En la primera etapa se seleccionaron 30 de las 76 zonas ($n = 30$), con un diseño de probabilidad proporcional al tamaño según la cantidad de viviendas. Por lo que la probabilidad de ser seleccionada una zona o manzana se explica en la fórmula 1:

$$P(k \in s) = \pi_k = n \cdot \frac{\#\text{ viviendas en la manzana k}}{\sum_{i=1}^{76} \#\text{ viviendas en la manzana i}} \tag{1}$$

Obteniendo en la primer etapa las 30 zonas coloreadas de la “Villa del Cerro” que se ven en la figura 1

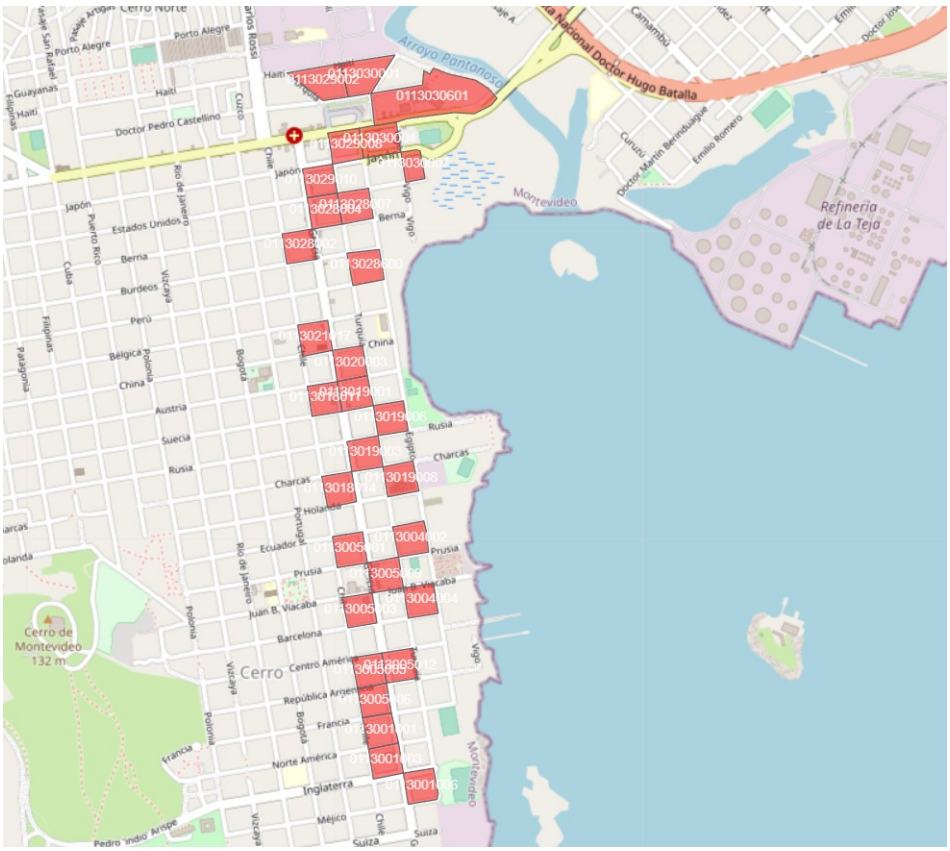

Figura 1: Zonas de ”Villa del Cerro”seleccionadas en la muestra

En la segunda etapa se hizo un diseño sistemático en cada una de las zonas, visitando 6 viviendas por zona. Para el diseño sistemático es necesario tener las viviendas ordenadas de alguna manera, y luego determinar un ’salto’ para obtener las casas seleccionadas en la muestra y así terminar con el tamaño de muestra deseado, en nuestro caso las viviendas están ordenadas al rededor de la manzana que constituye la zona.

Este diseño de muestreo es muy oportuno cuando se debe hacer trabajo de campo ya que el marco muestral puede presentar variaciones respecto a la realidad y no es posible determinar previamente los hogares exactos que se debe visitar, el encuestador va a la primer casa que sale sorteada, luego se saltea tantas casas como indique el ‘salto’ y encuesta a la siguiente, así hasta dar vuelta la manzana y obtener el tamaño de muestra deseado para la zona.

Este salto queda determinado por el tamaño de la zona (cantidad de viviendas $= N_h$) y el tamaño de muestra deseado dentro de cada zona (en nuestro caso $n_h = 6$). Por lo que el salto (r) será:

$$r_h = \frac{N_h}{n_h} \tag{2}$$

Y la probabilidad de pertenecer a la muestra de una vivienda en la segunda etapa, es decir, dado que su zona pertenece a la muestra de primera etapa es:

$$\pi_{l|k} = \frac{n_h}{N_h} \tag{3}$$

De esta forma, con las probabilidades de pertenecer a la primer y segunda etapa se tiene la probabilidad de pertenecer a la muestra de la vivienda l multiplicando ambas probabilidades, que utilizando la notación propuesta recién en la ecuación 1 queda:

$$\begin{aligned}\pi_l &= \pi_{l|k}\pi_k \\ &= \frac{n_h}{N_h} \times n \cdot \frac{\#\text{ viviendas en la manzana k}}{\sum_{i=1}^{76} \#\text{ viviendas en la manzana i}} \\ &= \frac{n_h}{N_h} \times n \cdot \frac{N_h}{\sum_{i=1}^{76} N_i} = \frac{n_h \cdot n}{\sum_{i=1}^{76} N_i}\end{aligned} \tag{4}$$

Luego, como el tamaño de muestra dentro de cada zona será $n_h = 6$, $n = 6$, y se tiene que el total de viviendas en nuestro universo de interés según el marco censal del año 2011 es 2507. La probabilidad de inclusión de la vivienda l es $\pi_l = \frac{6 \cdot 30}{2507} = \frac{180}{2507} = 0.0718$ para todas las viviendas y el ponderador original es:

$$w_l = \frac{1}{\pi_l} = \frac{1}{0.0718} = 13.928 \ , \forall l \tag{5}$$

2.2. Análisis Por No Respuesta

Se puede observar que yendo a 6 viviendas de cada una de las 30 zonas muestreadas en la primer etapa se debería tener un tamaño de muestra de 180 viviendas en "Villa del Cerro", pero al ir a encuestar un hogar no siempre se obtienen respuestas. Por lo tanto, se deben ajustarlos ponderadores originales para mitigar el sesgo que la no respuesta genera en las estimciones y para poder expandir las estimaciones a la población.

Se utilizará el enfoque estocástico, el cual, supone que cada vivienda tiene una cierta probabilidad de responder a la encuesta dado que pertenece a la muestra, entonces se tiene que la probabilidad de responder de una vivienda es:

$$P(R_l | l \in s) = \phi_l \tag{6}$$

Nuestro objetivo es crear un modelo para poder determinar las probabilidades de responder de cada una de las viviendas. Algunos modelos determinan esta probabilidad de responder a través de variables auxiliares como el sexo, la edad, si las personas dentro de la vivienda están ocupados o no, pero en este caso, dichas variables no son conocidas para los individuos que no responden. Por lo que debemos determinar la probabilidad de responder a través de otras variables.

Luego de realizada las encuestas se cuenta con 69 viviendas respondentes, es decir, con un 38.3 % de respuestas. Una de las opciones para ajustar por no respuesta es dividir todos los ponderadores entre 0.383 y seguiríamos obteniendo ponderadores iguales para todos los hogares, pero asignar esta probabilidad de respuesta a cada hogar sería inexacto

En la figura 2 se puede ver por una escala de color la cantidad de viviendas que respondieron de cada zona, y también se puede observar que hay indicios de que la no respuesta se comporta distinto a través de tres grandes grupos que están delimitados por las líneas rojas.

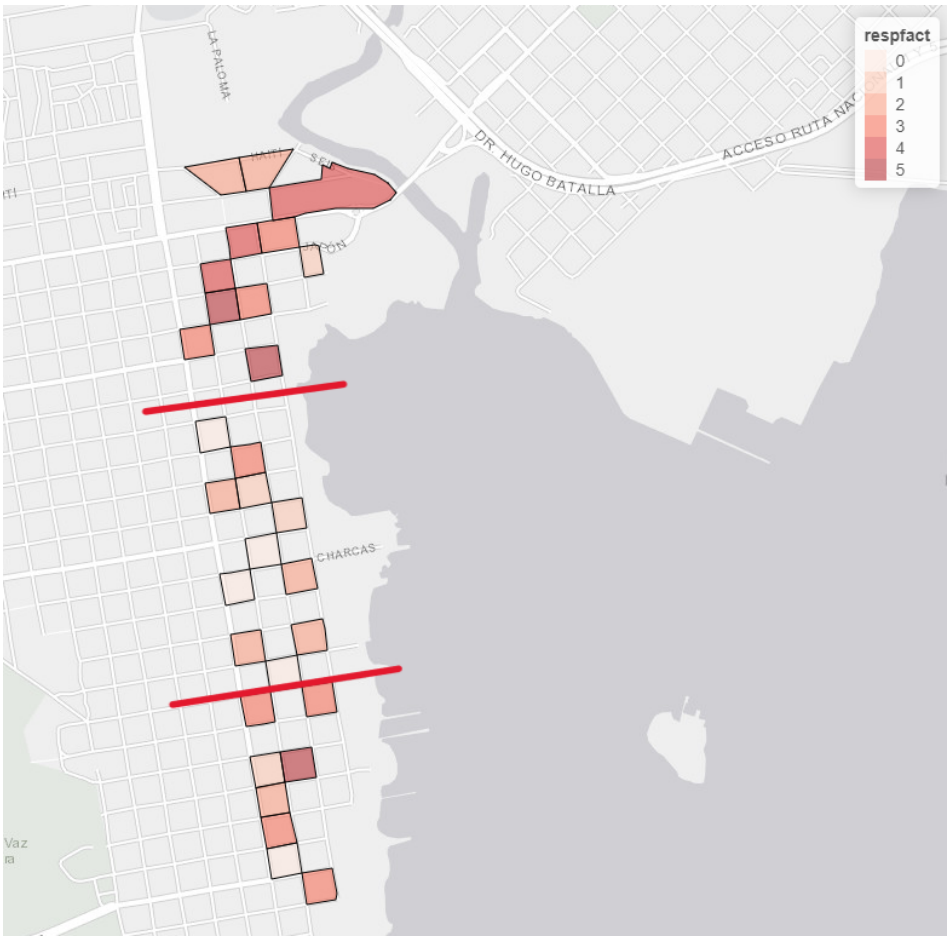

Figura 2: Zonas de "Villa del Cerro"seleccionadas en la muestra y coloreadas por cantidad de hogares respondentes

Las rectas del mapa buscan agrupar nuestros conglomerados (las zonas) en post-estratos cada uno con un comportamiento distinto frente a la no respuesta, y calcularemos las probabilidades de responder para cada post-estrato por separado, esta división en post-estrato corresponde a los distintos comportamientos de respuesta que se ve.

La probabilida de responder en cada post-estrato será la cantidad de respuesta obtenidas entre el total de muestra esperado dentro de él (es decir, el tamaño de muestra asignado). Por lo que si numeramos los post-estratos del 1 al 3 de Norte a Sur se tiene:

Post Estrato	N° Zonas	n° muestra	Respuestas	Prob Resp
1	11	66	36	0.55
2	11	66	14	0.21
3	8	48	21	0.44

Cuadro 1: Resumen de respuestas por post-estrato para el análisis de no respuesta.

Luego, los ponderadores ajustados por no respuesta de cada vivienda será:

$$w_l^{NR} = w_l \cdot \frac{1}{P(R_l)} \ , \forall l \tag{7}$$

En donde $P(R_l)$ es la probabildiad de responder de la vivienda l y queda determinada por el post-estrato al que pertenece su zona. Y como los ponderadores originales eran iguales para todas las viviendas, los ponderadores ajustados quedan:

$$\begin{aligned} w_l^{NR} &= 13.928 \cdot \frac{1}{0.55} = 25.53 \text{ para el post-estrato 1} \\ w_l^{NR} &= 13.928 \cdot \frac{1}{0.21} = 65.66 \text{ para el post-estreato 2} \\ w_l^{NR} &= 13.928 \cdot \frac{1}{0.42} = 31.83 \text{ para el post-estrato 3} \end{aligned} \tag{8}$$

Por último, al sumar los ponderadores de todas las viviendas se obtiene el total de viviendas de nuestra población que es 2507.

En cada etapa en que se ajustan los ponderadores respecto a los originales se debe calcular el efecto de Kish de los nuevos ponderadores, este efecto, llamado *deffK* se calcula a partir de la ecuación 9:

$$deffK(w) = 1 + cv(W) \tag{9}$$

En donde cv es el coeficiente de variación aplicado al vector de los ponderadores (w). Este valor se espera que tome valores menores a 1.5, de lo contrario se indica que los ponderadores fueron ajustados de manera excesiva. Para el caso de los ajustes por no respuesta, se tiene que $deffK(w^{NR}) = 1.19$ por lo que se puede continuar con la calibración.

3. Calibración

La última parte del ajuste de los ponderadores corresponde a la calibración, para poder obtener estimaciones con menor varianza.

La calibración de los ponderadores se realiza a partir de variables auxiliares presentes en el formulario y que a su vez son conocidas a nivel de los totales poblacionales gracias al censo 2011. La técnica de calibración, consiste en modificar los ponderadores con los que venimos trabajando en busca de alcanzar estos totales poblacionales cuando se expande la muestra.

Además, con la calibración de los ponderadores se logra disminuir el desvío de las estimaciones.

Para nuestra calibración se usará el sexo de las personas que residen en las viviendas, y el grupo de edad al que pertenecen según la clasificación utilizada en la encuesta[1].

Para determinar los totales poblacionales de cada una de las categorías de estas variables se cuenta con los datos del censo 2011, el tema de esto es que los datos del censo a nivel de personas no están detallados por zona, si no por segmento, y nuestra población objetivo está formada por segmentos pero algunos de estos segmentos no están enteros dentro de la población objetivo, como se puede ver en la figura 3, algunos segmentos quedan cortados por el límite de nuestra población objetivo.

[1] menos de 5, 5 a 12, 13 a 18, 19 a 24, 25 a 40, 41 a 60, 61 o más

Figura 3: Zonas y Segmentos de "Villa del Cerro"que son población objetivo

Entonces, los totales poblacionales por grupo de edad y sexo se obtuvieron con los totales poblacionales de los segmentos que pertenecen (al menos en alguna de sus zonas) a nuestra población objetivo, y se ajustó por la cantidad de personas que efectivamente viven en nuestra población objetivo según el censo 2011. Se tiene que en las zonas de nuestra población objetivo viven 6645 personas, y en los segmentos de los cuales se obtendrá la información del sexo y edad de todas las personas viven 9055. Por lo que los totales poblacionales de las categorías deben ajustarse multiplicándolos por el factor $6645/9055 = 0.734$, donde se asume que el sexo y la edad de las personas se distribuyen de manera uniforme dentro del segmento.

Además, como se contaba con datos faltantes del sexo o de la edad para algunas respuestas de la

Grupo	totales x grupo en segmentos	totales x grupo ajustado
menos de 5	694	510
5 a 12	905	664
13 a 18	778	571
19 a 24	707	519
25 a 40	2006	1472
41 a 60	2197	1612
más de 61	1768	1297

Cuadro 2: Tabla de totales poblacionales por grupo de edad en los segmentos de interés y ajustados por tamaño de población

Sexo	totales en segmento	totales ajustado
Hombre	4246	3116
Mujer	4809	3529

Cuadro 3: Tabla de totales poblacionales por sexo en los segmentos de interés y ajustados por tamaño de población

encuesta se decidió asignarles un valor aleatorio con probabilidad igual a las proporciones obtenidas en la muestra.

3.1. Calibración Raking

Como se comentó antes, la idea es ajustar nuestros ponderadores de forma que estimen exactamente las cantidades poblacionales de sexo y edad (que si bien en este caso fueron estimadas al hacer el ajuste porque no se contaba con la información a nivel de zona, se asumiran que son el valor real). A su vez, se buscará que los ponderadores cambien lo menos posible respecto a sus valores originales (los ajustados por no respuesta), así los estimadores son aproximadamente insesgados ya que los ponderadores de los que partimos producen estimadores insesgados. Este cambio mínimo se obtiene minimizando una función de distancia que mide la suma de las variaciones de los ponderadores, que bien puede ser la distancia de mínimos cuadrados entre los ponderadores originales y los nuevos, pero nosotros usaremos la distancia que produce el método de Raking.

acá entre parentesis no debería decir los ponderadores del diseño?

estimadores o estimaciones?

estimadores está bien, es una funcion

$$L(w^*, w^{NR}) = \sum_{i \in s^R} (w_i log(w_i^*/w_i^{NR}) - w_i^* - w_i) \tag{10}$$

En donde en la ecuación 10 se tiene que w^{NR} es el vector de ponderadores ajustado por No Respuesta, que son los que calibraremos ahora y w^* es el vector de ponderadores ajustado.

Un punto a tener en cuenta es que la encuesta se realizó a nivel de viviendas, y en cada vivienda tenemos sexo y edad de los individuos que viven en ella. Esta es la información que utilizaremos para calibrar los ponderadores, por lo que debemos desagregar las respuetas a nivel de individuo, y tomar la encuesta para esta sección como si estuviese relevada por congolomerados, donde cada conglomerado es un hogar, y se junta la información de todos los individuos que viven en él. Este paso es muy importante, porque luego tendremos que hacer que cada vivienda tenga un solo ponderador, por lo que se utilizará el promedio de los ponderadores de los individuos que viven en la vivienda.

Cuando se desagrega la encuesta por personas y se suman los ponderadores nos encontramos conque estos no suman la cantidad de personas de las personas en la población objetivo (6645), si no que esta suma da 6813, el método de Raking ajustará los ponderadores para lograr que estos repliquen los totales poblacionales por grupo de edad y por sexo, por lo que también deberá ajustar para que sumen

6645, por lo que los ponderadores de las personas tenderán a bajar un poco, más allá de los ajustes que sufran para replicar los totales poblacionales. Más adelante veremos los efectos de esto

El método de Raking hace referencia también a lo conocido como post-estratificación incompleta. La post-estratificación completa se utiliza cuando se desea "cruzar" las variables de control, por ejemplo en nuestro caso, si utilizaramos los hombres y mujeres dentro de cada grupo de edad en particular, exigiéndole a los ponderadores que estimen exactamente más categorías, en este caso 14 (2 categorías de sexo para cada uno de los 7 grupos de edad). Mientras que el raking, o post-estratificación incompleta genera ponderadores que estiman exactamente los totales de cada grupo etario y además los totales de cada sexo pero no predice los cruces de sexo y edad, es decir que por ejemplo no predice la cantidad exacta de hombres entre 25 y 40 años.

Este método lo que hace es calibrar los ponderadores haciendo una post-estratificación completa, pero de una variable a la vez, es decir, calibra los ponderadores para predecir exactamente los grupos de edad, luego, a estos ponderadores calibrados los vuelve a ajustar para estimar exactamente la cantidad de personas por sexo a nivel poblacional, y luego vuelve al punto anterior, y así hasta que el ajuste de los ponderadores entre un paso y otro sea mínimo y se logre una estimación exacta de todas las categorías de ambas variables. Entonces se tendrá que el ponderador del individuo i será:

$$w_i^* = g_i \cdot w_i^{NR} \tag{11}$$

En donde g_i es el factor de ajuste que se obtiene en el proceso de calibración para cada individuo. Al final, el ponderador del hogar será el promedio de los ponderadores de los individuos.

$$w_l^* = \frac{\sum_{i \in l} w_i^*}{\#\text{personas en el hogar } l} \tag{12}$$

Como se mencionó anteriormente, los ponderadores a nivel de personas seráin ajustados "hacia abajo"para ajustar a la cantidad de personas que se tiene en nuestro universo, por lo que ahora, cuando se promedian los ponderadores de los individuos dentro de los hogares y se suman veremos que estos no ajustan a la cantidad de hogares que teníamos inicialmente en nuestra muestra, ahora la suma de los ponderadores calibrados de los hogares da 2348, que es un 6 % por debajo de las 2507 viviendas que hay en la población objetivo, pero a su vez está más cerca de la cantida de viviendas particulares ocupadas que tenemos en el marco, que son 2252.

Para la calibración con el método de Raking, se utilizó la función *rake* de la librería *survey*, esta función recibe como argumento un diseño, en este caso hecho con *svydesign* de la misma librería, el nombre de las variables con la que se calibrará en nuestra muestra y una lista con los totales poblaciones. Además, se eligió esta función por encima de la función *calibrate* porque se entendió que se tiene más control sobre los totales poblaciones en el argumento correspondiente dentro de la función.

Dentro de *rake*, se le puede agregar un argumento respecto al control de la calibración, cómo el error máximo aceptado en las estimaciones poblacionales, las iteraciones máximas en el proceso de calibración o el valor máximo y mínimo de los factoes de ajuste.

Los dos primeros puntos suelen ajustarse cuando se obtiene algún error en la calibración con los valores que tiene por defecto. Mientras que el último punto se utiliza para mantener los ponderadores calibrados lo más cercano posible a los ponderadores bases y que las estimaciones sean aproximadamente insesgadas, además también se evitan ponderadores influyentes para algunas observaciones, o hasta ponderadores negativos en casos extremos. Para determinar si es necesario este ajuste debemos ver la distribución de los factores de ajuste en la figura 4, donde además se tiene que el factor de ajuste mínimo es 0.69 y el máximo es 1.47 en los hogares.

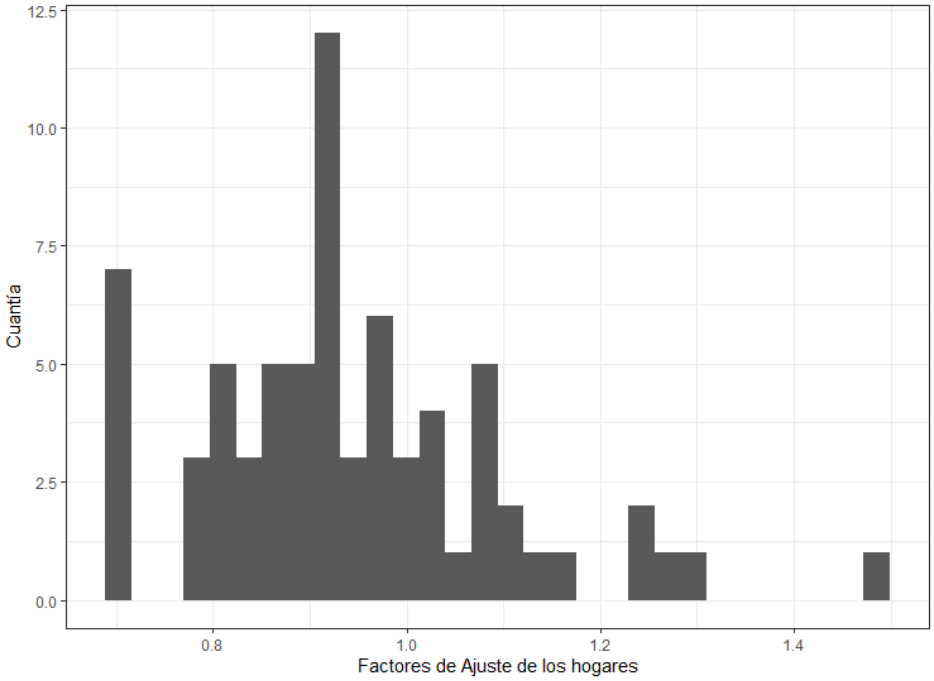

Figura 4: Histograma de los factores de ajuste

4. Estimaciones

Se realizan estimaciones sobre la cantidad de viviendas con hábito de reciclaje y sobre el conocimiento de las viviendas acerca de la cooperativa "La Paloma.[en] las viviendas de "Villa del Cerro". Para esto se tuvieron en cuenta distintos posibles diseños. En primer lugar se consideró un diseño por conglomerado por manzanas, y luego otro diseño en el que además se consideran los post-estratos creados en la etapa de ajuste por no respuesta.

Los distintos diseños generaran las mismas estimaciones, pero tendrán diferencia en el desvío de estas y el efecto diseño Deff que indica la eficiencia del diseño respecto al diseño simple [2], esto se puede ver en la tabla 4, donde se tiene que los primeros desvíos y deff presentados son para el diseño que no considera los post-estratos, mientras que los segundos son para el diseño que se asume estratificado en los post-estratos.

[2] $Deff(p) = V(p)/V(SI)$ donde p es el diseño utilizado y SI es el diseño simple, este valor nos indica si el diseño elegido es más o menos eficiente que el diseño simple (Si $Deff(p) < 1$ es más eficiente por tener menos varianza)

Respuesta	Porcentaje estimado	Desvío	Deff	Desvío 2	Deff 2
¿En el hogar recicla?					
Si	55.6 %	0.087	2.22	0.081	1.93
No	44.4 %	0.087	2.22	0.081	1.93
¿Conoce la cooperativa La Paloma?					
Si	24 %	0.056	1.24	0.052	1.07
No	77 %	0.056	1.24	0.052	1.07

Cuadro 4: Estimaciones de los porcentajes de viviendas que reciclan y conocen a la Cooperativa La Paloma en la Villa del Cerro

Se puede observar que cuando se estima con el diseño estratificado el desvío es levemente menor para las dos variable, esto se debe a que la varianza dentro de los estratos es un poco menor que la varianza total, y el diseño sin estratificar no tiene en cuenta esto. Es decir, el comportamiento de las viviendas es parecido dentro de un mismo estrato. Por lo tanto, a la hora de calcular la varianza de las estimaciones, es posible descomponerla en la suma de las varianzas dentro de cada estratos, lo cual genera una disminución en la varianza de las estimaciones y un mejor efecto diseño respecto a cuando no se estratifica.

También se estima los totales poblacionales en la tabla 5 para estas variables, en donde se puede observar que la suma de los hogares es 2348 y no 2507 según lo comentado en la etapa de calibración

Respuesta	Total estimado	Desvío	Deff	Desvío 2	Deff 2
¿En el hogar recicla?					
Si	1305	275.34	4.02	255.19	3.45
No	1043	194.45	2.00	175.59	1.63
¿Conoce la cooperativa La Paloma?					
Si	569	152.81	1.67	141.37	1.43
No	1780	194.41	2.69	171.20	2.09

Cuadro 5: Estimaciones de los totales

Se puede ver cómo son las estimaciones por estrato, tanto sobre los hábitos de reciclaje dentro de la viviendas, cómo del conocimiento de la cooperativa.

post estrato	Recicla	No Recicla	desvío(Recicla)	desvío(No Recicla)
1	0.52	0.48	0.09	0.09
2	0.60	0.40	0.18	0.18
3	0.54	0.46	0.13	0.13
post estrato	**Conoce La Paloma (Si)**	**No Conoce (No)**	**desvío(Si)**	**desvío(No)**
1	0.22	0.78	0.06	0.06
2	0.29	0.71	0.12	0.12
3	0.21	0.79	0.08	0.08

Cuadro 6: Estimaciones por post-estrato

En la tabla 6 se puede observar que el post-estrato 2 es en el que mayor proporción de hogares recicla, y también es en el que más conocimiento hay sobre la existencia de la cooperativa de trabajo “La Paloma”.