UNIVERSIDAD DE LA REPÚBLICA

Facultad de Ciencias Económicas y de Administración

Licenciatura en Estadística

Informe de Pasantía

Estimación en dominios de indicadores socioeconómicos a partir de la Encuesta Continua de Hogares

Mauricio Pittamiglio

Tutores:

Ignacio Alvarez-Castro

Juan José Goyeneche

Montevideo, Fecha.

Índice general

Índice de figuras

Índice de tablas

Capítulo 1

Introducción

El objetivo de este trabajo es generar una herramienta computacional que facilite el procesamiento de la Encuesta Continua de Hogares (ECH) mediante el cálculo de diversos indicadores y sus respectivos intervalos de confianza, aportando además una visualización interactiva de los resultados.

Como experiencia personal trabajé con la ECH, principalmente con el cálculo de indicadores a partir de esta encuesta, y contar con una herramienta como la realizada en este trabajo simplifica mucho la tarea.

Otra de las ventajas de esta aplicación es la disponibilidad en la web y el no necesitar de conocimientos en R o en algún otro programa estadístico para poder obtener las estimaciones puntuales e intervalos de confianza para los indicadores, pudiendo visualizar interactivamente los resultados a partir de tablas o distintos gráficos.

Las encuestas por muestreo no solo son utilizadas para obtener estimaciones a nivel de toda la población, sino que también pueden ser útiles para realizar estimaciones en subconjuntos de la población. Estos subconjuntos son denominados dominios, y los mismos pueden estar definidos por áreas geográficas, grupos demográficos u otro tipo de subpoblaciones.

Se trabajará con distintos indicadores socioeconómicos que se obtienen a partir de

estimaciones provenientes de la Encuesta Continua de Hogares. Los indicadores abordan distintas dimensiones, como ser, educación, salud, mercado laboral, entre otras. Estos se calculan tanto a nivel de toda la población (total país), así como para distintas aperturas (dominios o áreas de estimación), las cuales son definidas a nivel geográfico (departamentos). Algunos de estos indicadores son un insumo clave para la economía de los países y por ende para la toma de decisiones de las políticas públicas.

Capítulo 2

Métodos estadísticos

2.1. Introducción

Todos los indicadores presentados en este trabajo son estimaciones, por lo tanto se encuentran sujetas a errores de muestreo por el hecho de ser obtenidas a partir de una parte de la población (muestra) y no a partir de un censo. Estos errores de muestreo dependen de varios factores como ser: la variabilidad de la variable de interés, el tipo de indicador (parámetro), el dominio de estimación, la estrategia de selección de la muestra (diseño muestral) y los ajustes realizados para la construcción de los ponderadores.

El error de muestreo de la estimación de un parámetro es generalmente cuantificado por medio del error estándar de la estimación, que mide la variación de las estimaciones entre las distintas muestras y es el insumo principal para la construcción de intervalos de confianza.

Existen también los errores no muestrales, estos no se deben a la utilización de una muestra, sino a otras causas, como errores en la recolección o en el procesamiento de los datos, entre otros. Estos no serán abordados en este trabajo en particular.

En su mayoría los indicadores presentados son totales y ratios (tasas, promedios y

proporciones).

2.1.1. Dominios

En la práctica, las encuestas nacionales como la ECH, no son solamente utilizadas para hacer estimaciones a nivel de toda la población sino que también se realizan para diferentes subpoblaciones. Estas subpoblaciones para las cuales se computan de manera específica, tanto estimaciones puntuales como intervalos de confianza, son denominadas dominios. (SARNDAL)

En (CITAR ANDRES GUTIERREZ) se introduce la siguiente definición de dominios.

Definición 2.1 *Un dominio U_d es una sub-población específica o subgrupo poblacional que cumple las siguientes condiciones:*

1. *$U_d \subset U$, tal que $U = \bigcup_{d=1}^{D} U_d$*
2. *Si $k \in U_l$, entonces $k \notin U_d$ para $d \neq l$*
3. *El número de elementos en el dominio U_d es N_d y es llamado tamaño absoluto del dominio.*
4. *La proporción de elementos en el dominio U_d con respecto al tamaño poblacional es $P_d = \frac{N_d}{N}$ y se conoce como tamaño relativo del dominio.*

Los dominios pueden clasificarse como planeados o no planeados. Los dominios planeados son aquellas subpoblaciones para las cuales se quiere brindar estimaciones con precisiones específicas y los mismos son definidos como estratos de diseño, lo cual permite definir tamaños de muestra específicos. Mientras que los dominios no planeados, son aquellos que no fueron tenidos en cuenta a la hora de diseñar la muestra, debido a que generalmente no se tiene información en el marco de muestreo. Estos dominios son los más comunes en la práctica y un claro ejemplo de los mismos son aquellos definidos por las personas que se encuentran comprendidas dentro de

un tramo de edad específico.

2.2. Estimaciones puntuales

En toda encuesta por muestreo, el objetivo es seleccionar una muestra aleatoria s de tamaño n de una población U de tamaño N. Utilizando la muestra aleatoria s se busca obtener estimaciones de distintos parámetros de interés de la población U. Un parámetro de interés puede ser el total de la variable de interés y, el cual viene dado por:

$$t = \sum_{i=1}^{N} y_i = \sum_{i \in U} y_i \tag{2.1}$$

La mayoría de los indicadores calculados en este trabajo se obtienen a partir del cociente (ratio) entre totales de dos variables y y z, quedando definido de la siguiente manera:

$$R = \frac{\sum_{i \in U} y_i}{\sum_{i \in U} z_i} \tag{2.2}$$

Como se mencionó anteriormente las unidades, tanto hogares como personas, tienen una probabilidad de inclusión que viene determinada por el diseño muestral utilizado, la cual se denota como:

$$\pi_i = \text{Prob}(i \in s) \tag{2.3}$$

Posteriormente, la probabilidad de inclusión π_i de una unidad i en la muestra s es utilizada como insumo principal para calcular los ponderadores. El ponderador w de una unidad (hogar o persona) incluida en la muestra se define como el inverso de la probabilidad de selección π_i, es decir:

$$w_i = \frac{1}{\pi_i} \tag{2.4}$$

El ponderador w_i indica cuántas unidades de la población U que no fueron seleccionadas representa la unidad i que fue encuestada, es decir que pertenece a la muestra s.

Para poder computar las estimaciones de los parámetros de interés, se define el estimador Horvitz-Thompson.

$$\hat{t}_\pi = \sum_s \frac{y_i}{\pi_i} \tag{2.5}$$

El mismo también puede ser planteado de la siguiente forma:

$$\hat{t}_\pi = \sum_s w_i \times y_i \tag{2.6}$$

Este estimador es insesgado para el total $t = \sum_U y_i$. (CITAR SARNDAL POR LA DEMOSTRACIÓN). Por otra parte, el estimador de ratio o cociente R entre dos variables queda definido:

$$\hat{R} = \frac{\sum_{i \in s} w_i \times y_i}{\sum_{i \in s} w_i \times z_i} \tag{2.7}$$

El estimador $\hat{R}$ es un estimador no lineal ya que es un cociente entre dos estimadores, por lo tanto es aproximadamente insesgado. En toda encuesta suelen existir problemas de no respuesta y/o cobertura, es decir, el marco de muestreo donde físicamente es seleccionada la muestra no tiene un enlace perfecto con la población. Esto implica que en la práctica no se utilizen de forma directa los ponderadores para realizar las expansiones, sino que estos son ajustados por no respuesta y posteriormente calibrados a información conocida de la población. Estos nuevos ponderadores ajustados son los que finalmente son utilizados para la producción de las estimaciones de los

distintos indicadores (parámetros) y son los que se encuentran incluidos en las bases de datos públicas de la ECH. (CITAR INFORME OPP?)

Generalmente, los ajustes por no respuesta tienden a aumentar el error estándar (SE) de las estimaciones, ya que buscan únicamente reducir el posible sesgo ocasionado por las unidades que no respondieron a la encuesta, mientras que por otro lado, la calibración busca reducir los SE por medio de variables auxiliares o de control que se encuentran correlacionadas con las variables de interés de la encuesta.

2.3. Estimación de los errores estándar (SE)

Como la muestra es aleatoria, los estimadores son variables aleatorias y están, por lo tanto, sujetos a fluctuación. Esta fluctuación puede ser medida a partir del error estándar de la estimación. El SE es una medida de la precisión de la estimación y la misma puede ser utilizada para realizar inferencia de la población de interés por medio de la construcción de intervalos de confianza. El SE cuantifica la variación entre las estimaciones obtenidas entre las distintas muestras. Son calculados con el objetivo de añadir a las estimaciones puntuales una medida de calidad de las mismas. Por lo tanto, el problema se encuentra en poder obtener una estimación del SE que refleje, en lo posible, todas las fuentes de variabilidad del estimador o al menos las más importantes. Las fuentes de variabilidad vienen dadas por el diseño muestral implementado, así como también los distintos ajustes que son llevados a cabo para obtener los ponderadores finales w_i.

En la práctica muchos usuarios que utilizan la bases de datos de encuestas públicas para realizar sus propias inferencias no siempre tienen en cuenta todas las fuentes de variabilidad que afectan al error estándar. Esto, puede derivar en la realización de inferencias inválidas sobre los parámetros de la población. Uno de los motivos por los cuales surge este problema es el no contar con los insumos necesarios en la bases de datos públicas de las encuestas sobre el diseño de muestreo. En el caso

de la ECH, para los años anteriores a 2018 no se cuenta con la disponibilidad de información clave sobre el diseño muestral como los son las unidades primarias de muestreo (UPM) y los estratos a los que pertenece cada individuo.

Para un diseño complejo como el de la ECH, no existen métodos exactos para calcular los SE, lo que implica que se deba recurrir a distintas estrategias o métodos que aproximan de mejor forma la variabilidad de las estimaciones. Dentro de la amplia gama de métodos que existen se optó por el método del último conglomerado para computar los SE, que es el más usado en la práctica y se encuentra implementado en el paquete *survey* del *R*. Este método puede manejar tanto parámetros lineales como no lineales. (INFORME TAMBIEN)

2.3.1. Método del último conglomerado

El método del último conglomerado es utilizado ampliamente en la práctica para la estimación de los SE en diseños por conglomerados en una o varias etapas de selección, en donde, los conglomerados (UPM) son seleccionados con probabilidades proporcional al tamaño.

A continuación se desarrolla el método del último conglomerado aplicado a un muestreo aleatorio, estratificado, por conglomerados y en dos etapas de selección, como lo es el de la ECH. Pero previamente, se introduce la notación y conceptos necesarios para desarrollar dicho método.

La población de interés U es particionada en grupos o estratos $U_1, U_2, ..., U_h, ..., U_H$ excluyentes. En cada uno de los estratos, se selecciona una muestra aleatoria s_h de forma independiente en dos etapas de selección. Sea M_h la cantidad de UPM en el estrato h y N_h la cantidad de unidades en dicho estrato. En una primera etapa, dentro de un estrato h se selecciona m_h UPM entre las M_h bajo un muestreo aleatorio sin reposición con probabilidad proporcional al tamaño en base a una medida de tamaño (MOS). La probabilidad de selección de la UPM j perteneciente al estrato

h viene dada por:

$$\pi_{jh} = m_h \times \frac{\text{MOS}_j}{\sum_{j=1}^{M_h} \text{MOS}_j}$$

Luego, dentro de cada UPM j seleccionada en la primera etapa dentro del estrato de diseño h, se seleccionan n_{jh} unidades, generalmente, con igual probabilidad de selección bajo un muestreo aleatorio simple o sistemático con arranque aleatorio. La probabilidad de selección de la unidad i dentro de la UPM j viene dada por:

$$\pi_{i|jh} = \frac{n_{jh}}{\text{MOS}_j}$$

La probabilidad de selección final de la unidad en la muestra queda definida de la siguiente forma:

$$\pi_{ijh} = \pi_{jh} \times \pi_{i|jh}$$

Por lo que el peso muestral de la unidad i queda definido como el inverso de la probabilidad de selección, es decir:

$$w_i = \pi_{ijh}^{-1}$$

En la ECH, las UPM corresponden a conglomerados de zonas censales, mientras la medida de tamaño MOS para la definición de las probabilidades de inclusión, es el número de viviendas N_j de la UPM j según datos del último censo de población disponible (2011 en este caso). Finalmente, el número de viviendas en cada UPM seleccionada en la primera etapa es fijo, es decir $n_j = \alpha$. Bajo los dos criterios anteriores, las probabilidades de selección y por tanto los pesos muestrales son iguales dentro de un mismo estrato:

$$\pi_{ijh} = \pi_{jh} \times \pi_{i|jh} = m_h \times \frac{N_j}{\sum_{j=1}^{M_h} N_j} = \frac{m_h \times \alpha}{N_h} = \frac{n}{N}$$

Una vez definidos los ponderadores w_i, se pueden computar las estimaciones de los distintos indicadores. Sin embargo, el problema es que bajo este tipo de diseños no existen estimaciones insesgadas de los errores estándar, ya que el insumo principal para las estimaciones de los SE utilizando el estimador Horvitz-Thompson, son las probabilidades de inclusión de segundo orden, es decir, la probabilidad de que dos unidades cualquiera i y k pertenezcan a la muestra.

El método del último conglomerado asume que la mayor variabilidad en la estimación proviene de la primera etapa de muestreo, es decir, en la selección de las UPMs. Por lo tanto, la estimación del SE cuadrado (varianza) utilizando el método del último conglomerado para la estimación del total de una variable cualquiera y para un muestreo aleatorio, estratificado, por conglomerados y en varia etapas de selección viene dada por:

$$\widehat{\mathrm{SE}}^2(\hat{t}) = \hat{V}(\hat{t}) = \sum_{h=1}^{H} \frac{1}{m_h(m_h-1)} \times \sum_{j \in s_h} (\hat{t}_{jh} \times m_h - \hat{t}_h)^2 \tag{2.8}$$

donde:

m_h es la cantidad de UPM pertenecientes al estrato h seleccionadas en la muestra.

$\hat{t}_{jh} = \sum_{s_{jh}} w_i \times y_i$ es la estimación del total de la variable y en la UPM j perteneciente al estrato h.

$\hat{t}_h = \sum_{h=1}^{H} \hat{t}_{jh}$ es la estimación del total de la variable y en el estrato h.

Por otra parte, si el objetivo es estimar el total de la variable y pero para un dominio o área de estimación cualquiera d, el SE se obtiene reemplazando en la ecuación anterior **??** la variable y por la variable y_d la cual, toma el valor y para los casos de la muestra incluidos en el dominio d y cero en otro caso. Esta técnica se utiliza

para añadir una variabilidad extra a las estimaciones, producto de que el tamaño de muestra en el dominio (por ejemplo, personas comprendidas en un determinado tramo de edad y dentro de un determinado departamento) es aleatorio.

2.3.2. Linearización de Taylor

La linearización de Taylor es un método ampliamente utilizado para aproximar SE de parámetros no lineales (como ser un ratio). La idea es aproximar un estimador no lineal por medio de una función lineal. Una vez realizado lo anterior y teniendo en cuenta el diseño muestral utilizado se obtiene una aproximación del SE del estimador. Para el caso del estimador de una tasa o ratio $R = t_y/t_z$, es aproximado utilizando el desarrollo de taylor de primer orden y dicha aproximación queda definida como:

$$\hat{R} \approx R + \frac{1}{t_z} \times \sum\nolimits_{i \in s} w_i \times (y_i - R z_i)$$

Luego, el estimador de la varianza (o error estándar al cuadrado) se obtiene reemplazando la variable y por una variable $r_i = y_i - \hat{R} z_i$ y añadiendo el término $1/\hat{t}_z^2$.

$$\widehat{\mathrm{SE}}^2(\hat{R}) = \hat{V}(\hat{R}) = \frac{1}{\hat{t}_z^2} \times \sum_{h=1}^{H} \frac{1}{m_h(m_h - 1)} \times \sum_{j \in s_h} (\hat{t}_{r,jh} \times m_h - \hat{t}_{r,h})^2 \tag{2.9}$$

donde:

$\hat{t}_z = \sum_{i \in s} w_i \times z_i$ es la estimación del total de la variable z.

$\hat{t}_{r,jh} = \sum_{s_{jh}} w_i \times r_i$ es la estimación del total de la variable r en la UPM j perteneciente al estrato h.

$\hat{t}_{r,h} = \sum_{h=1}^{H} \hat{t}_{r,jh}$ es la estimación del total de la variable r en el estrato h.

Al igual que para el caso de un total, el SE de la estimación de un ratio R para un dominio d se obtiene reemplazando en la ecuación **??** las variables y y z por sus

correspondientes variables extendidas en el dominio, y_d y z_d.

2.4. Intervalos de confianza

A partir de la estimación de los errores estándar, se obtienen los intervalos de confianza. Estos son una medida de calidad de las estimaciones y son computados con el fin de poder extraer conclusiones sobre la población.

Los intervalos de confianza se suelen calcular de la siguiente manera:

$$\hat{\theta} \pm \text{constante} \times \widehat{\text{SE}}(\hat{\theta})$$

Donde el estimador $\hat{\theta}$ se asume tiene una distribución aproximadamente normal. Por lo que el intervalo de confianza para un parámetro cualquiera θ con un nivel de confianza $100(1-\alpha))\,\%$ queda definido como:

$$\hat{\theta} \pm z_{1-\alpha/2} \times \widehat{\text{SE}}(\hat{\theta})$$

Particularmente se trabajará con una confianza del 95 %, por lo que al sustituir el valor de la constante:

$$\hat{\theta} \pm 1{,}96 \times \widehat{\text{SE}}(\hat{\theta}) \tag{2.10}$$

Un intervalo con nivel de confianza del 95 % significa que, en promedio, de cada 100 muestras obtenidas bajo el mismo diseño, el intervalo contiene al verdadero valor del parámetro en 95 de ellas.

Capítulo 3

Indicadores

3.1. ECH

La Encuesta Continua de Hogares (ECH) es una encuesta realizada de forma ininterrumpida por el Instituto Nacional de Estadística (INE) desde el año 1968. La ECH es un tema de interés nacional, ya que brinda los indicadores oficiales del mercado laboral y de ingresos de los hogares y las personas con periodicidad mensual, trimestral, semestral y anual, entre otros.

El diseño muestral de la ECH es aleatorio, estratificado, por conglomerados y en dos etapas de selección. Como primer paso para la selección de la muestra la población es particionada en estratos. El primer nivel es geográfico y corresponde a los diecinueve departamentos del país y a la zona metropolitana. En un segundo nivel, cada una de las localidades dentro del departamento es clasificada en cuatro categorías: i) localidades con más de 20.000 habitantes, ii) localidades entre 5.000 y menos de 20.000 habitantes, iii) localidades entre 200 y 5.000 habitantes y iv) áreas rurales y localidades con menos de 200 habitantes. Para Maldonado y Rocha se agrega un nivel más correspondiente a las zonas balnearias. En el departamento de Montevideo y zona metropolitana se conforman cinco y tres estratos socioeconómicos respectivamente.

Dentro de cada estrato y de forma independiente se selecciona, en una primera etapa, conglomerados de zonas censales (Unidades Primarias de muestreo - UPM) bajo un muestreo aleatorio, sistemático y proporcional al tamaño (PPS), utilizando como medida de tamaño (MOS) la cantidad de viviendas particulares según datos provenientes del último censo de población y vivienda del año 2011. Luego, en segunda etapa se seleccionan cinco viviendas titulares y dos viviendas suplentes en cada una de las UPM seleccionadas en la primera etapa, con igual probabilidad de selección bajo un muestreo aleatorio simple.

Dentro de cada vivienda seleccionada se relevan características propias de la vivienda así como del hogar que reside en la misma e información de todos los integrantes que conforman el hogar. Esto implica que la ECH brinde información para computar distintos indicadores no solo a nivel de hogar sino también a nivel de persona. Es por esto, que el INE publica dos bases de datos, una base con información a nivel del hogar y la vivienda y otra base con información a nivel de personas.

El tamaño de muestra efectivo anual de la ECH, es decir, la cantidad de hogares que cumplen los criterios de ser elegibles y respondieron se encuentra en el entorno de 42000 hogares.

Los datos de la muestra son ponderados, de forma de obtener estimaciones tanto a nivel nacional, departamental y de otros dominios de estudio, entre ellos, sexo y tramos de edades. Existen varias etapas para la determinación de los ponderadores finales de la ECH. El componente principal es el inverso de la probabilidad de selección del hogar y sus integrantes en la muestra de la ECH, denominado ponderador original. El ponderador original de un hogar perteneciente a un estrato cualquiera se define como el inverso de la tasa de muestreo en el estrato.

Los ponderadores originales son ajustados en varias etapas una vez concluida la recolección de la información. El primer ajuste se realiza en base a la no respuesta obtenida en campo, es decir, teniendo en cuenta las encuestas efectivamente realizadas. Un segundo ajuste se efectúa utilizando técnicas de calibración en base a las

proyecciones de la población residente en viviendas particulares. Esto implica que la ECH expanda a la población proyectada para el año de referencia para cada uno de los departamentos del país, y para la estructura de la población a nivel del total del país por sexo y cinco tramos de edades (0 a 14 años, 15 a 29 años, 30 a 49 años, 50 a 64 años y 65 años o más). (CITAR INFOME DEL INE)

3.2. Listado de indicadores

Los indicadores presentados en este trabajo son calculados todos los años por el Observatorio Territorio Uruguay (OTU) perteneciente a la Oficina de Planeamiento y Presupuesto (OPP), y se encuentran disponibles en su página oficial. Estos indicadores se dividen en 7 categorías:

- Educación
- Salud
- Mercado laboral
- Ingresos y bienestar
- Tecnología y comunicación
- Demografía
- Viviendas y hogares

Dentro de la categoría *Educación* se encuentran 17 indicadores, referidos a tasas de asistencia a distintos niveles educativos, tasas de analfabetismo, promedio de años de educación, entre otros.

En la categoría *Salud* hay 3 indicadores que evalúan la afiliación a emergencias móviles y el tipo de atención en salud.

Hay 13 indicadores disponibles dentro de *Mercado laboral*, además de los principales indicadores de este rubro (tasas de empleo, desempleo, actividad), se encuentran

otros referidos a la informalidad, categoría de ocupación, entre otros.

La categoría *Ingresos y bienestar* cuenta con 8 indicadores, índices de pobreza (en personas y hogares) y algunos indicadores presentes también en la categoría *Mercado laboral* (empleo, desempleo e informalidad).

En *Tecnología y comunicación* se encuentran 7 indicadores referidos a tenencia de distintos dispositivos como celulares o computadoras y a la utilización de estos dispositivos y de internet.

Demografía contiene 3 indicadores referidos a residencias previas o lugar de nacimiento de las personas.

Finalmente la categoría *Viviendas y hogares* cuenta con 8 indicadores, todos ellos referidos a distintas características de los hogares.

Recordando que hay indicadores que se encuentran en varias categorías, en total se trabaja con 53 indicadores socio-económicos, desagregando todos por departamento y a su vez hay algunos que se desagregan también por sexo o tramos etarios.

En anexo (REFERENCIA) se presenta un listado completo de los indicadores trabajados.

Capítulo 4

Computación estadística

4.1. Estimación de los indicadores

Para replicar el diseño muestral de la ECH y poder llevar a cabo el computo de los indicadores, se utiliza el software libre *R*, más precisamente el paquete *srvyr*. Este paquete utiliza las funciones del paquete *survey* antes mencionado, permitiendo el uso de un estilo de sintaxis inspirado en el paquete *dplyr*. (CITAR LOS PAQUETES)

Este paquete *srvyr* cuenta con funciones tales como *survey_mean* para calcular medias o proporciones, *survey_ratio* para ratios o *survey_total* para totales. Estas funciones utilizan el valor que toma el indicador en cada uno de los individuos (hogares o personas según corresponda) de la muestra y los ponderadores, dando como resultado tanto las estimaciones puntuales que correspondan como las estimaciones de los errores estándar, utilizando las ecuaciones y metodologías descritas en este documento para realizar los cálculos.

A continuación se presenta un breve ejemplo de como se utilizan las funciones de esta librería.

```
> ind<-diseno%>%
+       filter(eval(parse(text=filtro)))%>%
```

```
+       group_by(dpto)%>%
+       summarise(estimacion.tot= survey_mean(eval(parse(text=variable))))
```

Estas líneas de código son la base de las funciones creadas para el cálculo de los indicadores, donde en este caso se le indica el objeto de diseño, se le aplica el filtro necesario, se agrupa por departamento y se estima la media de la variable indicada para cada departamento, dando como resultado un objeto (data frame) con las estimaciones puntuales y errores estándar para el indicador deseado.

Para la estimación de los indicadores se crearon 9 funciones que varían dependiendo de los individuos involucrados (personas y hogares), la forma en la que se calcula cada indicador y las desagregaciones necesarias. Todas estas funciones estiman el indicador tanto a nivel de cada departamento como de todo el país y tienen como salida una tabla (data frame) con los resultados.

Las primeras 7 funciones se utilizan para los indicadores que se obtienen a partir de la base de personas de la ECH, mientras las últimas 2 trabajan con la base de hogares.

La primera función se utiliza para los indicadores que necesitan del cálculo de la proporción de individuos que cumplen determinada condición (por ejemplo quienes tienen estudios terciarios), desagregando además por sexo.

La segunda función es utilizada para los indicadores en los cuales es necesario calcular la proporción de distintas categorías de una variable, y también se desagrega por sexo.

Las funciones 3, 4 y 5 son similares, calculan proporciones al igual que en la función 1, pero desagregando en tramos de edad en lugar de por sexo. La diferencia entre estas tres funciones es que desagregan en distintos tramos etarios.

La función 6 se utiliza unicamente para la estimación del indicador *Promedio de años de educación de las personas de 25 años y más por sexo*. La variable en este caso es cuantitativa, se promedia y se desagrega por sexo.

La función 7 se utiliza para los indicadores que hacen referencia a tasas brutas, es decir, necesitan del cálculo de un ratio.

Los indicadores *Población total por condición de actividad por sexo* y *Personas por tipo de atención en salud*, fueron calculados de manera particular, sin el uso de ninguna función.

Las funciones 8 y 9, como se mencionó anteriormente, son las utilizadas para la estimación de los indicadores que se obtienen a partir de la base de hogares de la ECH. La función 8 se utiliza para los indicadores que requieren el cálculo de la proporción de hogares que cumplen con determinada condición, mientras que en la 9, la variable en cuestión presenta más de 2 categorías.

El indicador *Hogares por tipo de relación con la vivienda* fue estimado de manera particular, sin el uso de ninguna función.

Todas las funciones necesitan de algunos insumos para estimar los indicadores correspondientes, estos son:

- El diseño muestral.
- El filtro a aplicar (de ser necesario). Este determina la población objetivo para la cual el indicador será calculado.
- Variable(s) de interés: variable(s) necesaria(s) para el cálculo del indicador.
- Vector con categorías de la variable de interés (en caso de ser necesario).

Se crean además funciones para el cálculo de los intervalos de confianza de las estimaciones, funcionando de la misma manera que las funciones mencionadas anteriormente pero obteniendo una salida diferente, 9 de ellas dan como salida una tabla con los extremos inferiores de los intervalos, mientras las otras 9 hacen lo mismo con los límites superiores.

A continuación se presenta un ejemplo del código para la aplicación de una de las funciones. El indicador es *Utilización de computadora el último mes por sexo.*

```
> funcion1(pr,"e27>5","e61==1")
```

Este indicador es estimado en base a personas y con desagregación por departamento y por sexo. Se utiliza la *funcion1*, el diseño utilizado es *pr*, el filtro es $e27 > 5$ (edad mayor a 5 años) y la variable utilizada es $e61 == 1$ la cual es una variable presente en la ECH que hace referencia a la utilización de computadora en el último mes.

4.1.1. Descripción de los diseños

El insumo *diseño muestral* que utilizan las funciones creadas para el cálculo de los indicadores depende de la información disponible con la que cuente el usuario. En las bases públicas del INE no se cuenta con información del estrato ni de la UPM a la que pertenece el individuo, por lo que en la App aparecen dos opciones. La opción *pública* se utiliza cuando no se cuenta con esta información, mientras que en caso de contar con esta se utiliza la opción *estratos+UPM*.

En caso de no contar con la información de estratos y UPM, los cálculos de los intervalos de confianza son incorrectos, ya que esta información es esencial para poder realizar una aproximación correcta de los SE y así poder construir los intervalos de confianza.

Cabe aclarar que INE hizo pública esta información para los años 2018 y 2019.

4.2. Shiny App

Shiny es una librería que permite crear fácilmente aplicaciones web interactivas a partir de R. La interactividad de estas aplicaciones permite la manipulación de los datos sin la necesidad de manipular el código, es decir, en caso de estar disponible en alguna página web, puede ser fácilmente utilizada por personas que no tengan conocimientos de R.

Estas aplicaciones cuentan con programación reactiva. Esto significa que al cambiar los valores de entrada (inputs) se volveran a ejecutar las partes del código de R correspondientes, lo que a su vez hará que se actualicen las salidas (outputs) modificadas, es decir, la modificación de un valor reactivo llevará automaticamente a todas las expresiones reactivas que dependen directa o indirectamente de este valor a ejecutarse nuevamente.

Se cuenta además con widgets pre-construidos que facilitan la construcción de aplicaciones bonitas e interactivas. Un widget es un elemento web que los usuarios pueden interactuar con él enviando mensajes a la aplicación Shiny.

Una Shiny app contiene al menos dos componentes: un script para la interfaz del usuario (ui) donde se controla el diseño de la aplicación; y un script para realizar los cálculos necesarios (server). En el server se debe especificar como convertir los valores de entrada (inputs) en resultados (outputs).

4.2.1. Esta app

Esta aplicación procesa los datos de la ECH, particularmente se realizó en base a la encuesta del año 2019, pero también funciona perfectamente para los años (..)2015 a 2018 y está pensada para que funcione también para los próximos años, en caso de que la encuesta no tenga cambios abruptos en su formato y estructura. Para el cálculo de los indicadores se utilizan un grupo de variables de la ECH, lo que hace a la app dependiente de esas variables, es decir, en caso de que esas variables cambien de nombre o dejen de estar disponibles, la app debe ser actualizada. Tanto en la app como en los anexos (REFERENCIAR ANEXO) de este informe, estará disponible un listado con las variables utilizadas.

En esta aplicación en particular, previo a la creación del ui y el server, se cargan las librerías a utilizar, los datos espaciales necesarios para los gráficos de mapas y se crean las funciones que se utilizan para el cálculo de los indicadores. Esta parte del

código se corre solo una vez al iniciar la app.

4.2.2. UI

La ui cuenta con tres partes, el encabezado (*dashboardHeader*), la barra lateral (*dashboardSidebar*) y el contenido (*dashboardBody*).

Dentro del encabezado solo aparece el título y un link que lleva directamente a la página del INE donde están disponibles las bases de datos de la ECH.

La barra lateral es utilizada para cargar las bases de datos y para seleccionar la sección que se desee visualizar, ya sea la pestaña de introducción o las correspondientes a las diferentes categorías de los indicadores. Para cargar los datos se utiliza la función *fileInput* y se deben cargar por separadas las bases de personas y de hogares de la ECH y el archivo que contiene la información acerca de los estratos y UPM (en caso de contar con el). También se incluye un widget mediante la función *selectInput* para especificar si las bases utilizadas son las públicas o si se cuenta con la información de estratos y UPM.

En el body se especifica el contenido de cada una de las secciones. La ventana introductoria cuenta con... .

La sección de los indicadores cuenta con 3 pestañas.

En la primera de estas pestañas se selecciona el indicador que se quiere calcular y se pueden visualizar dos tipos de gráficos para este indicador, un gráfico de mapa donde, en caso de que corresponda, se puede seleccionar la categoría a graficar, y un diagrama de barras (apiladas, agrupadas o simple según el caso). En la segunda pestaña se puede observar una tabla para las estimaciones puntuales del indicador seleccionado y un botón de descarga (*downloadButton*) para poder descargar dicha tabla en formato csv. Mientras la tercer pestaña cuenta con las tablas para los límites de los intervalos de confianza de las estimaciones, divididas en dos paneles, uno para el límite inferior y otro para el superior.

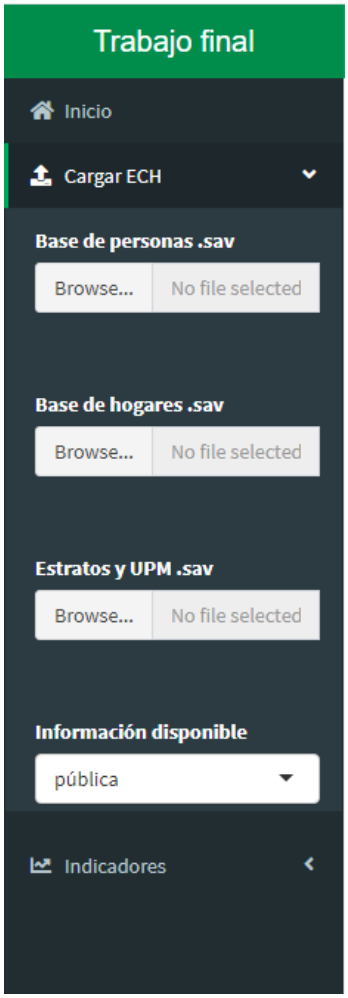

Figura 4.1: Barra lateral.

4.2.3. Server

En esta parte de la App se realizan todos los cálculos necesarios. En primer lugar se crean todos los elementos reactivos (se guardan las bases de datos cargadas, los diseños y todos los indicadores) utilizando la función *reactive*. Luego se crean todos los componentes de salida con funciones render, las tablas con *renderDT* y los gráficos con *renderPlotly*. Tanto las tablas como los gráficos se crean en outputs diferentes para cada una de las categorías de indicadores.

Plotly hace los gráficos interactivos, cuenta con varias opciones para poder visualizarlos de manera más precisa y permite descargarlos como imagen.

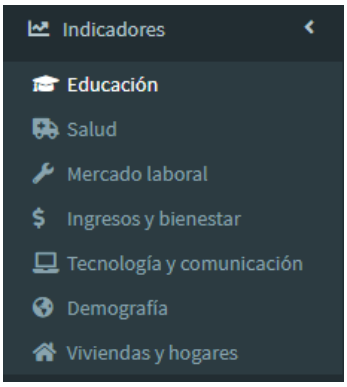

Figura 4.2: Categorías de indicadores.

Figura 4.3: Pestañas.

Capítulo 5

Ejemplo de aplicación

En esta sección se presenta el procedimiento aplicado y los resultados obtenidos para el caso de dos indicadores en particular, la `tasa de desempleo por sexo` y la `utilización de computadora el último mes por sexo`. Utilizando la ECH del 2019 y la información sobre los estratos y unidades primarias de muestreo (UPM).

En primer lugar se cargan las bases de datos necesarias, disponibles en la página web del INE. Primero se debe cargar la base de personas, luego la de hogares y en tercer lugar, el archivo con la información sobre estratos y UPM.

Una vez cargadas las bases se selecciona dentro de la sección de indicadores, la categoría *Mercado laboral* donde se encuentra este indicador. Una vez dentro de esta sección, se selecciona el indicador que se desea obtener.

En la misma pestaña se observan como resultados los dos tipos de gráficos antes mencionados, los cuales se presentan a continuación.

Como se puede observar en la imagen, al colocar el cursor sobre uno de los departamentos, el gráfico nos brinda el valor de la estimación para dicho departamento. Esto también es posible en el diagrama de barras anterior y es una de las ventajas del uso de gráficos interactivos creados con la librería *plotly*.

Para el gráfico de mapa se debe seleccionar la categoría a graficar, en este caso las

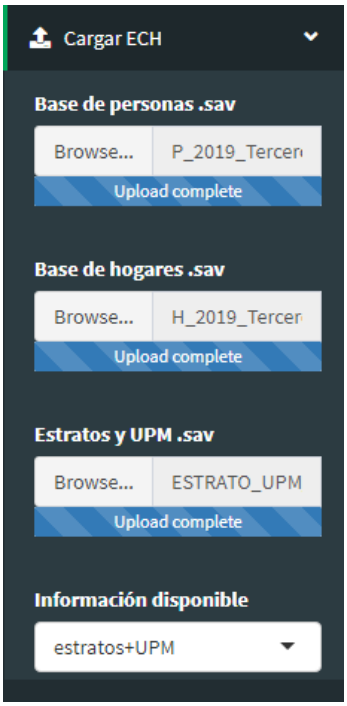

Figura 5.1: Bases de datos.

opciones son *Hombre*, *Mujer* o *Total*.

En la segunda pestaña de esta sección se puede observar una tabla como la presentada en la figura **??**, que contiene las estimaciones puntuales del indicador seleccionado, a nivel departamental y nacional.

Por último, en la pestaña *IC*, se observan tablas con los límites de los intervalos de confianza al 95 % de las estimaciones anteriores.

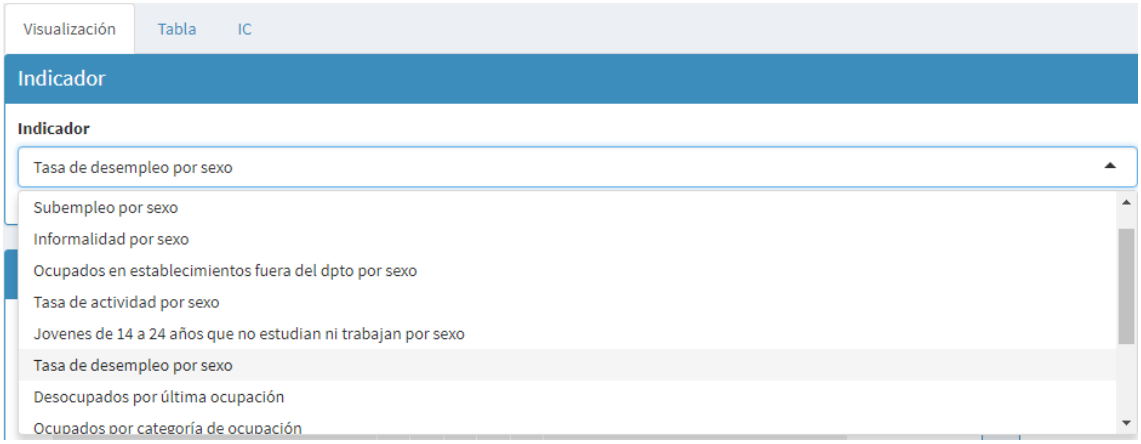

Figura 5.2: Selección del indicador.

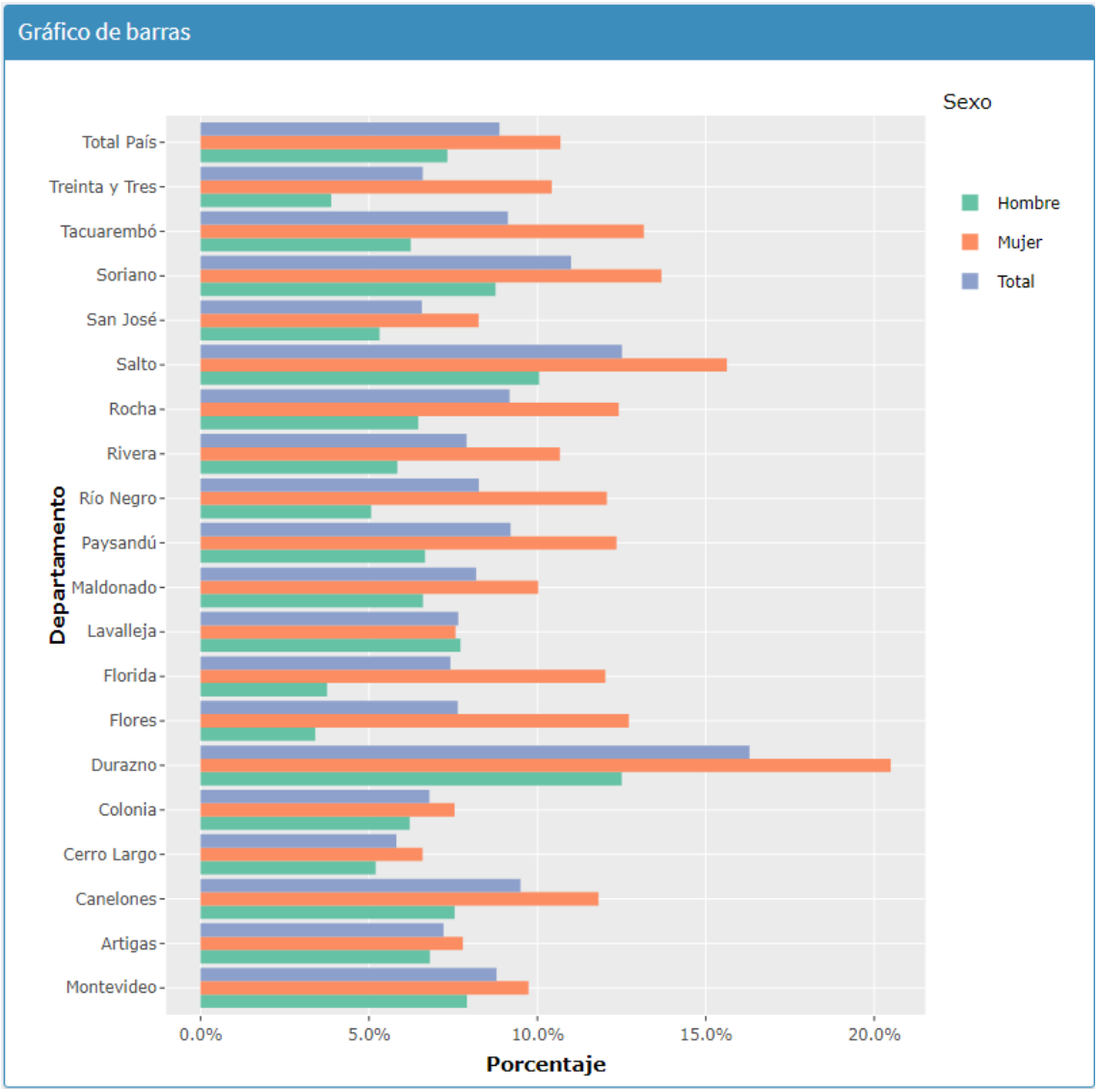

Figura 5.3: Gráfico de barras de la tasa de desempleo a nivel departamental y de todo el país desagregando por sexo.

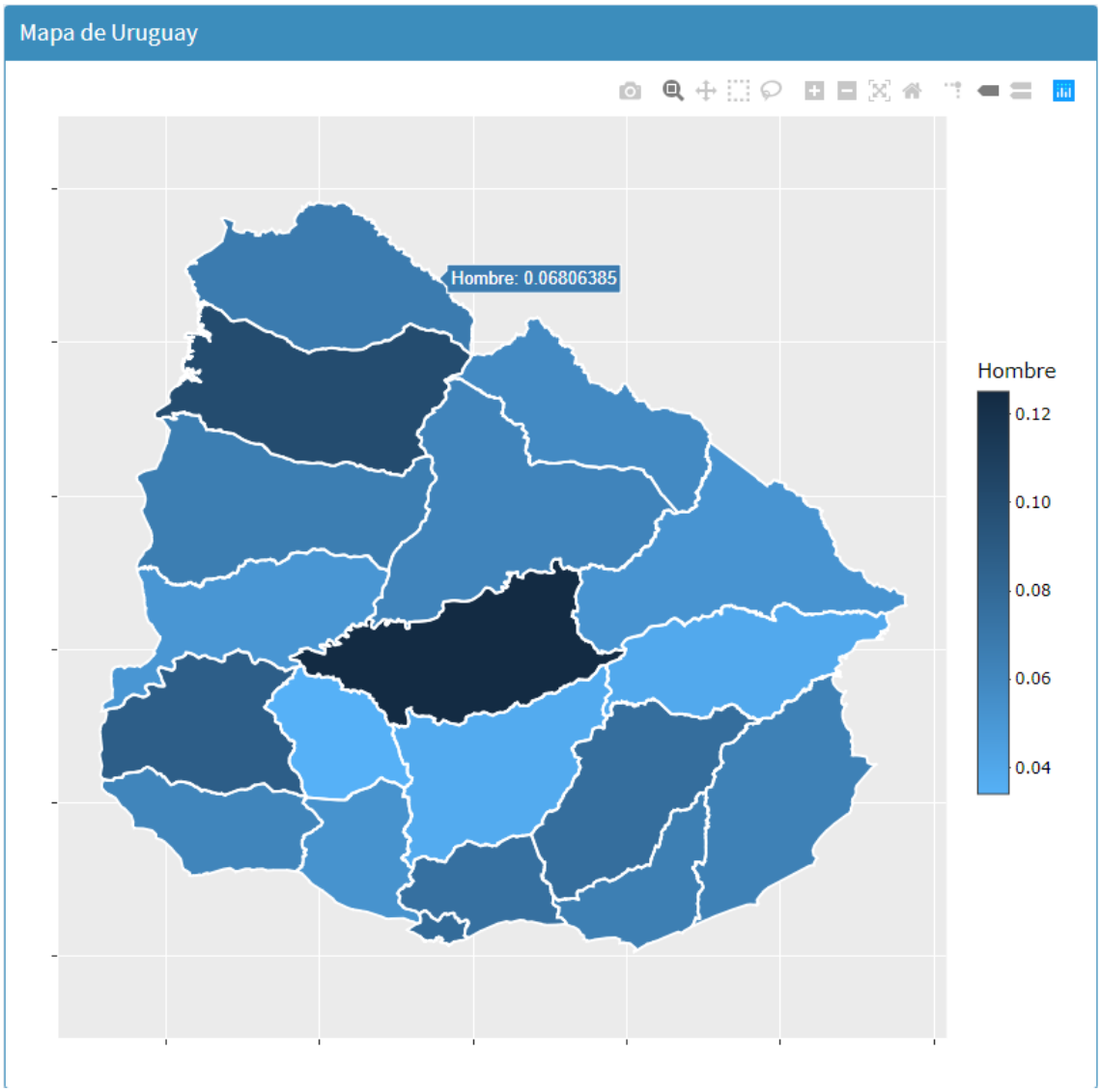

Figura 5.4: Tasa de desempleo a nivel departamental para la categoría Hombres.

Figura 5.5: Categorías para visualizar en el mapa.

Search:

depto	Hombre	Mujer	Total
Montevideo	0.0791	0.0974	0.0878
Artigas	0.0681	0.0779	0.0721
Canelones	0.0754	0.1181	0.0950
Cerro Largo	0.0520	0.0659	0.0581
Colonia	0.0621	0.0754	0.0679
Durazno	0.1250	0.2049	0.1630
Flores	0.0340	0.1271	0.0763
Florida	0.0375	0.1202	0.0742
Lavalleja	0.0771	0.0757	0.0765
Maldonado	0.0660	0.1002	0.0818
Paysandú	0.0666	0.1235	0.0920
Rio Negro	0.0506	0.1206	0.0826
Rivera	0.0584	0.1066	0.0789
Rocha	0.0646	0.1241	0.0917
Salto	0.1005	0.1562	0.1251
San José	0.0531	0.0825	0.0657
Soriano	0.0876	0.1368	0.1100
Tacuarembó	0.0624	0.1316	0.0912
Treinta y Tres	0.0388	0.1043	0.0659
Total País	0.0733	0.1068	0.0887

Descargar

Figura 5.6: Tabla con las estimaciones puntuales de la tasa de desempleo para los 19 departamentos y para todo el país en el año 2019.

Inferior | Superior

Search:

depto	Hombre	Mujer	Total
Montevideo	0.0734	0.0913	0.0836
Artigas	0.0468	0.0550	0.0566
Canelones	0.0672	0.1076	0.0883
Cerro Largo	0.0360	0.0448	0.0455
Colonia	0.0478	0.0577	0.0564
Durazno	0.0915	0.1666	0.1352
Flores	0.0099	0.0747	0.0497
Florida	0.0205	0.0864	0.0566
Lavalleja	0.0526	0.0489	0.0584
Maldonado	0.0525	0.0836	0.0709
Paysandú	0.0494	0.0962	0.0764
Río Negro	0.0303	0.0850	0.0635
Rivera	0.0426	0.0800	0.0638
Rocha	0.0479	0.0979	0.0762
Salto	0.0801	0.1302	0.1061
San José	0.0376	0.0627	0.0536
Soriano	0.0632	0.1082	0.0902
Tacuarembó	0.0445	0.1003	0.0743
Treinta y Tres	0.0209	0.0638	0.0445
Total País	0.0699	0.1027	0.0860

Figura 5.7: Tabla con los límites inferiores de los intervalos de confianza al 95 % de las estimaciones puntuales de la tasa de desempleo para los 19 departamentos y para todo el país en el año 2019.

Inferior | Superior

Search:

depto	Hombre	Mujer	Total
Montevideo	0.0847	0.1035	0.0921
Artigas	0.0893	0.1007	0.0877
Canelones	0.0837	0.1286	0.1016
Cerro Largo	0.0680	0.0870	0.0708
Colonia	0.0763	0.0931	0.0794
Durazno	0.1586	0.2432	0.1907
Flores	0.0582	0.1796	0.1029
Florida	0.0545	0.1540	0.0917
Lavalleja	0.1017	0.1025	0.0946
Maldonado	0.0795	0.1168	0.0927
Paysandú	0.0839	0.1507	0.1076
Río Negro	0.0709	0.1562	0.1018
Rivera	0.0742	0.1333	0.0941
Rocha	0.0813	0.1504	0.1072
Salto	0.1209	0.1823	0.1441
San José	0.0686	0.1024	0.0778
Soriano	0.1119	0.1654	0.1298
Tacuarembó	0.0803	0.1629	0.1082
Treinta y Tres	0.0567	0.1447	0.0874
Total País	0.0766	0.1110	0.0914

Figura 5.8: Tabla con los límites superiores de los intervalos de confianza al 95 % de las estimaciones puntuales de la tasa de desempleo para los 19 departamentos y para todo el país en el año 2019.

Capítulo 6

Discusión