Desagregación de Estimaciones en Áreas Pequeñas un enfoque bayesiano

Invalid Date


OBJETIVOS DE DESARROLLO SOSTENIBLE
1 FIN DE LA POBREZA
2 HAMBRE CERO
3 SALUD Y BIENESTAR
4 EDUCACIÓN DE CALIDAD
5 IGUALDAD DE GÉNERO
6 AGUA LIMPIA Y SANEAMIENTO
7 ENERGÍA ASEQUIBLE Y NO CONTAMINANTE
8 TRABAJO DECENTE Y CRECIMIENTO ECONÓMICO
9 INDUSTRIA, INNOVACIÓN E INFRAESTRUCTURA
10 REDUCCIÓN DE LAS DESIGUALDADES
11 CIUDADES Y COMUNIDADES SOSTENIBLES
12 PRODUCCIÓN Y CONSUMO RESPONSABLES
13 ACCIÓN POR EL CLIMA
14 VIDA SUBMARINA
15 VIDA DE ECOSISTEMAS TERRESTRES
16 PAZ, JUSTICIA E INSTITUCIONES SÓLIDAS
17 ALIANZAS PARA LOGRAR LOS OBJETIVOS
OBJETIVOS DE DESARROLLO SOSTENIBLE

Algunas metas del ODS2 (Hambre cero)

De aquí a 2030, poner fin al hambre y asegurar el acceso de todas las personas, en particular los pobres y las personas en situaciones de vulnerabilidad, incluidos los niños menores de 1 año, a una alimentación sana, nutritiva y suficiente durante todo el año.

- Prevalencia de la subalimentación.
- Prevalencia de la inseguridad alimentaria moderada o grave en la población, según la Escala de Experiencia de Inseguridad Alimentaria.

Algunas metas del ODS8 (Empleo decente)

De aquí a 2030, lograr el empleo pleno y productivo y el trabajo decente para todas las mujeres y los hombres, incluidos los jóvenes y las personas con discapacidad, así como la igualdad de remuneración por trabajo de igual valor.

- Tasa de desempleo, desglosada por sexo, edad y personas con discapacidad.

Principio fundamental de la desagregación de datos

Los indicadores de los Objetivos de Desarrollo Sostenible deberán desglosarse, siempre que sea pertinente, por ingreso, sexo, edad, raza, etnicidad, estado migratorio, discapacidad y ubicación geográfica, u otras características, de conformidad con los Principios Fundamentales de las Estadísticas Oficiales.

Resolución de la Asamblea General - 68/261

Limitaciones de las encuestas.

¿Qué es el coeficiente de variación?

El coeficiente de variación es una medida de error relativo a un estimador, se define como:

$$cve\left(\hat{\theta}\right) = \frac{se\left(\hat{\theta}\right)}{\hat{\theta}}$$

Muchas veces se expresa como un porcentaje, aunque no está acotado a la derecha, y por eso es conveniente a la hora de hablar de la precisión de una estadística que viene de una encuesta.

Estándares de alerta en algunos países (encuestas de hogares)

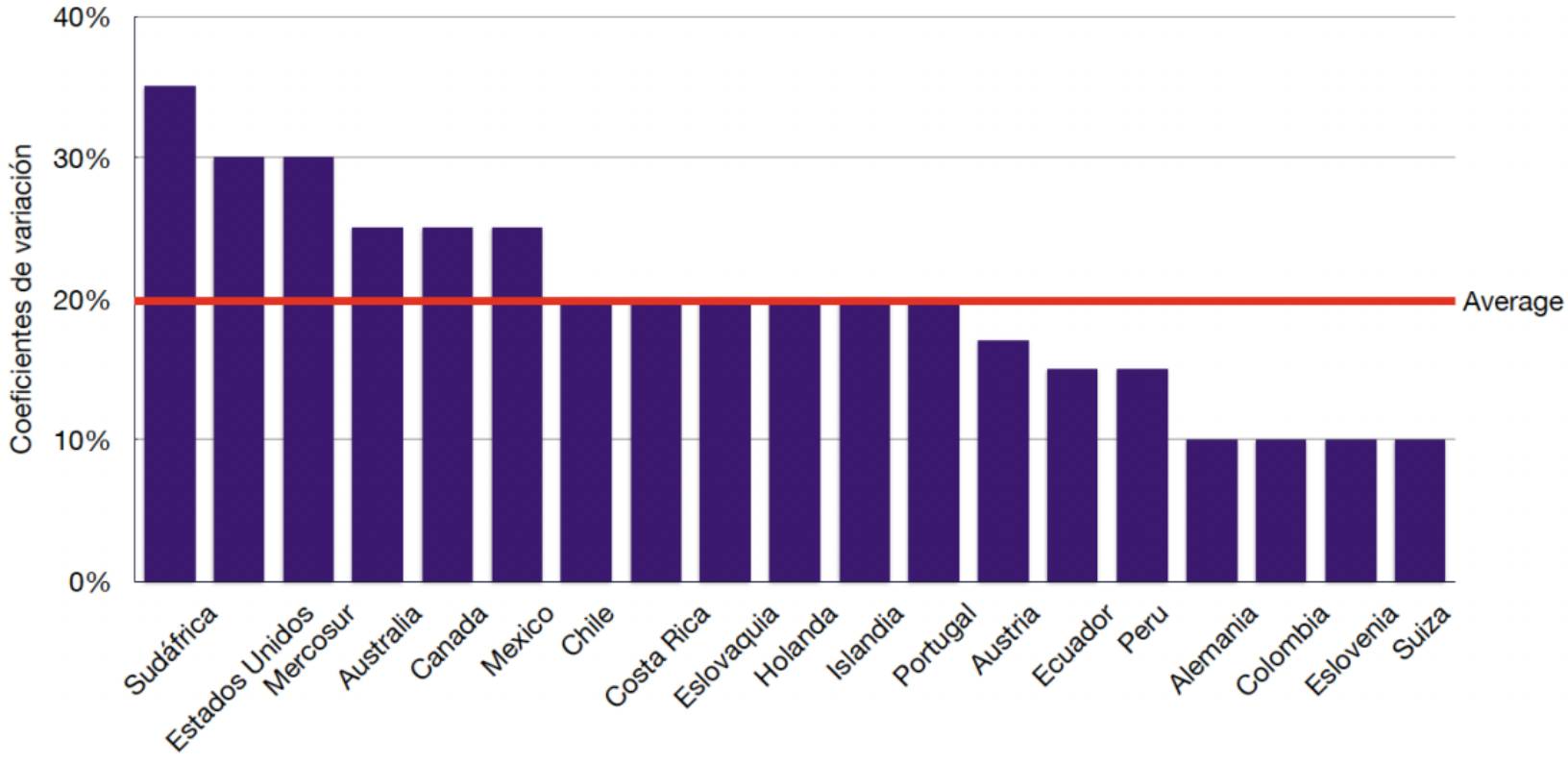

Figure 1: Alertas sobre los coeficientes de variación

Algunas alertas definidas en la publicación

Cuando se sobrepasa el umbral del coeficiente de variación aparecen algunas de las siguientes alertas:

- No se publica
- Usar con precaución.
- Las estimaciones requieren revisiones, no son precisas y se deben usar con precaución.
- Poco confiable, menos preciso.
- No cumple con los estándares de publicación.
- Con reserva, referencial, cuestionable.
- Valores muy aleatorios, estimación pobre.

Dominios de estudio y subpoblaciones de interés

Una encuesta se planea con el fin de generar información precisa y confiable en los dominios de estudio que se han predefinido. Sin embargo, existen subgrupos poblacionales que la encuesta no abordó en su diseño, y sobre los cuales se quisiera una mayor precisión.

- Incidencia de la pobreza desagregado por departamento o provincia (tamaño de muestra conocido y planificado).
- Tasa de desocupación desagregada por sexo (tamaño de muestra aleatorio, pero planificado).
- Tasa de asistencia neta estudiantil en primaria desagregada por quintiles de ingreso (tamaño de muestra aleatorio).

Precisión de los estimadores

Debido a que una encuesta es una investigación parcial sobre una población finita, es necesario saber que:

- A partir de una encuesta, no se calculan indicadores, sino que se estiman con ayuda de los datos de la encuesta.
- Es necesario calcular el grado de error que se comete al no poder realizar una investigación exhaustiva. Este error es conocido como el error de muestreo.
- La precisión de un estimador está supeditada al intervalo de confianza.

Entre más angosto sea el intervalo, más precisión se genera y por ende se tiene un menor error de muestreo.

El tamaño de muestra efectivo

- En las encuestas de hogares, con diseños de muestreo complejos, no existe una sucesión de variables que sean independientes e identicamente distribuidas.
- La muestra $y_1, \ldots, y_n$ no es un vector en el espacio n-dimensional, donde se asume que cada componente del vector puede variar por sí mismo.
- La dimensión final del vector $(y_1, \ldots, y_n)$ es mucho menor que n, puesto que existe una forma jerárquica en la selección de los hogares y a la interrelación de la variable de interés con las UPMs

El tamaño de muestra efectivo

El tamaño de muestra efectivo se define como sigue:

$$n_{efectivo} = \frac{n}{Deff}$$

En donde Deff es el efecto de diseño que depende de: *1.* El número de encuestas promedio que se realizaron en cada UPM. *2.* La correlación existente entre la variable de interés y las mismas UPMs.

Es posible considerar que, si el tamaño de muestra efectivo no es mayor a un umbral, entonces la cifra no debería ser considerada para publicación.

Grados de libertad

En las subpoblaciones los grados de libertad no se consideran fijos sino variables.

$$gl = \sum_{h=1}^{H} v_h \times (n_{Ih} - 1)$$

Note que ν_h es una variable indicadora que toma el valor uno si el estrato h contiene uno o mas casos de la subpoblación de interés, n_{Ih} es el número de UPMs en el estrato. En el caso más general, los grados de libertad se reducen a la siguiente expresión:

```
gl = #UPMs - #Estrato
```

Uso de métodos SAE

Justificación

- Los estimadores directos, basados solo en unidades de muestreo observadas para cada área pequeña, no son suficientemente confiables.
- Tamaño de muestra pequeño o incluso ninguna unidad observada (falta de información).
- El coeficiente de variación (CV) es demasiado alto para el indicador objetivo a nivel de área.

Incremento del coeficiente de variación

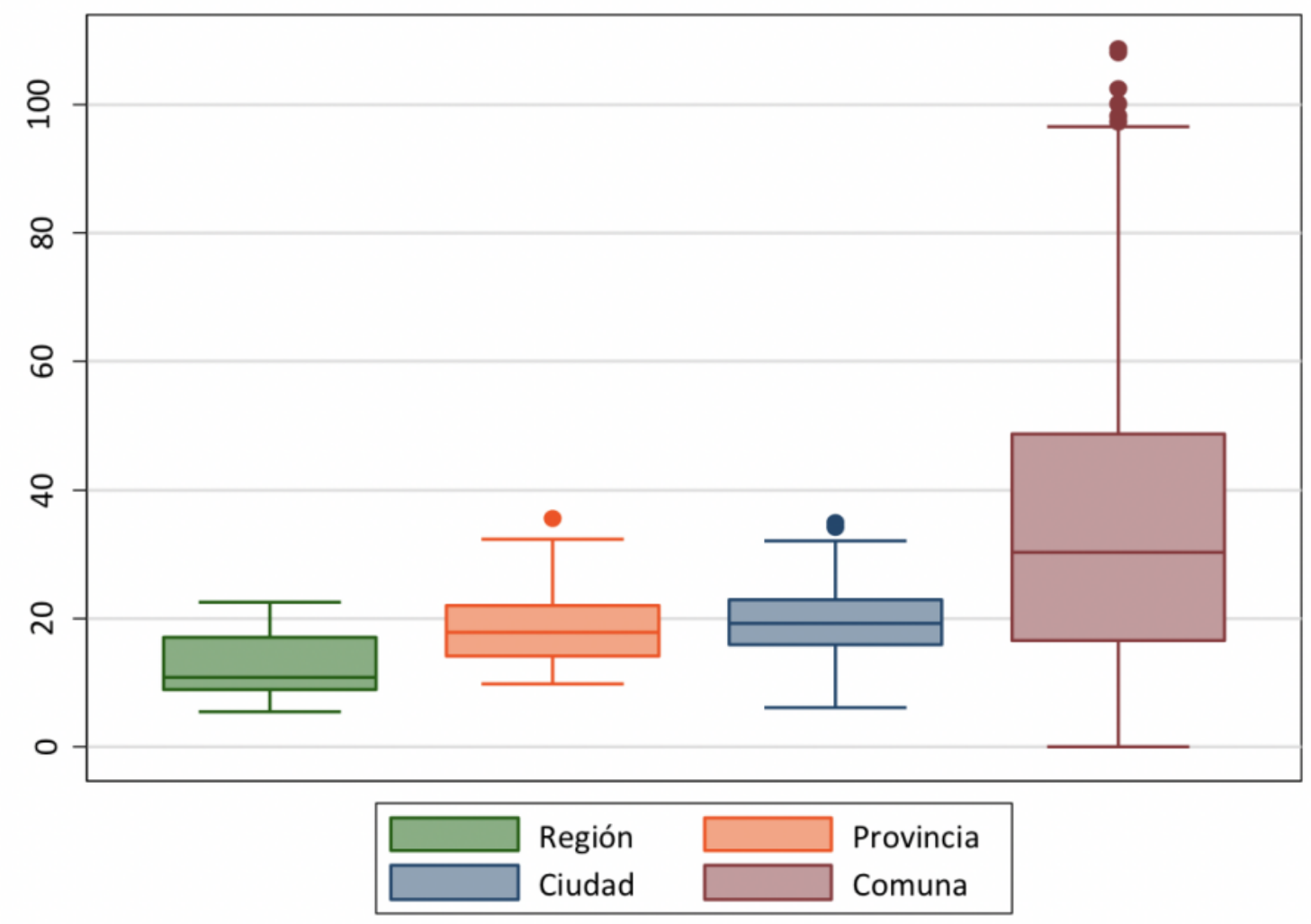

Figure 2: Distribución de los coeficientes de variación en Chile

Justificación

Cuando los estimadores directos no son confiables para algunos dominios de interés, existen dos opciones:

1 Sobremuestreo: aumentar el tamaño de la muestra en los dominios de interés (aumento de los costos).

2 Aplicar técnicas estadísticas que permitan estimaciones confiables en esos dominios, métodos SAE.

¿Qué es un área pequeña?

- La mayoría de las encuestas nacionales están planificadas para entregar estimaciones confiables a nivel nacional y regional pero a niveles más bajos se reduce la precisión.
- Un área pequeña es un dominio para el cual el tamaño de muestra específico no es suficientemente grande para obtener estimaciones confiables.
- Habitualmente son dominios no planificados y su tamaño de muestra esperado es aleatorio y es más grande a medida que aumenta el tamaño de la población del área.

¿Qué es un área pequeña?

La subpoblación de interés puede ser una zona geográfica o subgrupos socioeconómicos.

- Geográfico: provincias, áreas del mercado de trabajo, municipios, sectores censales para medir por ejemplo la tasa de desempleo a nivel comunal.
- Dominio de subgrupos específicos: edad $\times$ sexo $\times$ raza dentro del ámbito geográfico de una zona, para medir por ejemplo la tasa de desempleo por sexo o edad específica en las zonas urbanas.

Algunos métodos

- Los estimadores SAE se dividen en dos tipos principales dependiendo de cómo se aplican los modelos a los datos dentro de las áreas pequeñas: nivel de área y nivel de unidad.
- Los estimadores de área pequeña se basan en cálculos de nivel de área si los modelos vinculan la variable de interés y con variables auxiliares x específicas del área.

Algunos métodos

- Se llaman modelos a nivel de unidad si se vinculan valores individuales para las variables auxiliares específicas de la unidad.
- Los estimadores basados en áreas pequeñas se calculan a nivel de área si los datos de la unidad no están disponibles.
- También pueden ser calculados si los datos de nivel de unidad están disponibles resumiéndolos en el nivel de área apropiado.

Proceso de estimación

Figure 3: Producción de estadísticas con SAE

Consideraciones

- Todos los métodos SAE requieren datos auxiliares a nivel del área pequeña desde el cual toman prestada la fuerza.
- La efectividad de los métodos SAE depende del grado de asociación entre la variable de interés y los datos auxiliares.
- La búsqueda de buenas variables auxiliares es crítica, incluida la construcción imaginativa de tales variables.
- Los datos auxiliares deben medirse de manera consistente a través de las áreas pequeñas, pero pueden incluir estimaciones de muestras grandes con error de muestreo conocido.

Desafíos

- Aumento de las tasas de no respuesta.
- Aumento de costos, menos financiación.
- Aumento de la demanda de estimaciones para dominios pequeños como por raza, etnia o pobreza.
- Aumento de la demanda de estimaciones de áreas pequeñas.
- Aumento de la complejidad en los contenidos de los cuestionarios y por lo tanto la carga de respuesta.
- Aumento de la demanda de análisis secundarios, uso público y archivos de datos de uso restringido.

Función Generalizada de Varianza (FGV)

¿Cuál es la importancia de la Función Generalizada de Varianza?

- La varianza del estimador directo es un insumo crucial en el modelo de áreas.
- No es posible calcular la varianza del estimador directo a nivel de dominio.
- En dominios con un tamaño de muestra muy pequeño, las estimaciones de varianza pueden ser poco fiables.
- Se sugiere la utilidad de un modelo de suavizamiento de las varianzas.
- El propósito del suavizamiento es eliminar el ruido y la volatilidad en las estimaciones de varianza para obtener una señal más precisa del proceso.

La Función Generalizada de Varianza

Hidiroglou (2019) establece que: $E_{\mathcal{MP}}\left(\hat{\theta}_d^{dir}\right) = x_d^T\beta$ y $V_{\mathcal{MP}}\left(\hat{\theta}_d^{dir}\right) = \sigma_u^2 + \tilde{\sigma}_d^2$, en donde el subíndice $\mathcal{MP}$ hace referencia a la inferencia doble que se debe tener en cuenta en este tipo de ajustes.

- $\mathcal{M}$ hace referencia a la medida de probabilidad inducida por el modelamiento y la inclusión de las covariables auxiliares (x_d).
- $\mathcal{P}$ hace referencia a la medida de probabilidad inducida por el diseño de muestreo complejo que induce las estimaciones directas.

Estimación de la Varianza de Muestreo

La FGV consiste en ajustar un modelo log-lineal a la varianza directa estimada. Partiendo del hecho de que se tiene acceso a un estimador insesgado de σ^2, denotado por $\hat{\sigma}^2$ se tiene que:

$$E_{\mathcal{MP}}\left(\hat{\sigma}_d^2\right) = E_{\mathcal{M}}\left(E_{\mathcal{P}}\left(\hat{\sigma}_d^2\right)\right) = E_{\mathcal{M}}\left(\sigma_d^2\right) = \tilde{\sigma}_d^2$$

La anterior igualdad puede interpretarse como que un estimador insesgado y simple de $\tilde{\sigma}_d^2$ puede ser $\hat{\sigma}_d^2$.

Modelos de Suavizamiento

Rivest y Belmonte (2000) proponen modelos de suavizamiento para estimar las varianzas directas. Estos modelos se definen de la siguiente manera:

$$\log\left(\hat{\sigma}_d^2\right) = z_d^T \alpha + \varepsilon_d$$

En donde z_d es un vector de covariables explicativas que son funciones de x_d, α es un vector de parámetros que deben ser estimados, ε_d son errores aleatorios con media cero y varianza constante, que se asumen idénticamente distribuidos condicionalmente sobre z_d.

Estimación Suavizada

- La estimación suavizada de la varianza de muestreo está dada por:

$$\tilde{\sigma}_d^2 = E_{\mathcal{MP}}\left(\sigma_d^2\right) = \exp\left(z_d^T \alpha\right) \times \Delta$$

En donde, $E_{\mathcal{MP}}\left(\varepsilon_d\right) = \Delta$.

- Haciendo uso del método de los momentos, se tiene el siguiente estimador insesgado para Δ:

$$\hat{\Delta} = \frac{\sum_{d=1}^{D} \hat{\sigma}_d^2}{\sum_{d=1}^{D} \exp\left(z_d^T \alpha\right)}$$

Estimación de parámetros

- La estimación del coeficiente de parámetros de regresión está dada por la siguiente expresión:

$$\hat{\alpha} = \left(\sum_{d=1}^{D} z_d z_d^T \right)^{-1} \sum_{d=1}^{D} z_d \log\left(\hat{\sigma}_d^2\right)$$

- Y el estimador suavizado de la varianza muestral está definido por:

$$\hat{\tilde{\sigma}}_d^2 = \exp\left(z_d^T \hat{\alpha}\right) \hat{\Delta}$$

Datos: Gran Encuesta Integrada de Hogares (GEIH) de Colombia.

La Gran Encuesta Integrada de Hogares (GEIH) del 2018 en Colombia, utilizó un diseño muestral complejo que incluyó la estratificación de la población en zonas urbanas y rurales, junto con un muestreo por conglomerados. La muestra seleccionada fue significativa, permitiendo la recolección de datos de manera representativa en todo el país. En total, se utilizaron 98,000 Unidades Primarias de Muestreo (UPM) para tener estadísticas confiables a Nivel Nacional, Regiones Geográficas, Ciudades principales y Áreas Urbanas/Rurales, Estratos Socioeconómicos.

Set de datos

Table 1: GEIH Colombia

dam	dam2	wkx	upm	estrato	pobreza
05	05360	127.2	010126005360	051	0
05	05360	127.2	010126005360	051	0
05	05360	127.2	010126005360	051	0
05	05360	127.2	010126005360	051	0
05	05360	127.2	010126005360	051	0
05	05360	125.7	010126005360	051	1
05	05360	125.7	010126005360	051	1
05	05360	125.7	010126005360	051	1
05	05360	125.7	010126005360	051	1
05	05360	115.9	010126005360	051	1

Diseño muestral

Para definir el diseño muestral a partir de una base de datos de encuesta se usan las librerías `survey` y `srvyr`.

```
library(survey)
library(srvyr)
options(survey.lonely.psu = "adjust")

diseno <-
  as_survey_design(
    ids = upm,
    weights = wkx,
    strata = estrato,
    nest = TRUE,
    .data = encuesta
  )
```

Estimaciones directas por dominio

Para la estimación directa de la proporción se emplea la función `direct.supr`, disponible en el archivo `0Source_FH.R`. Está función realiza las estimaciones y criterios de calidad en una encuesta de muestreo complejo con diseño estratificado y por conglomerados.

```
directodam2 <- direct.supr(design.base = diseno,
                           variable = pobreza,
                           group = dam2,
                           upm = upm,
                           estrato = estrato)
```

Dominios seleccionados

- Mínimo 50 observaciones por dominio.
- Efecto de diseño (Deff) mayor a 1.
- Mínimo 3 grados de libertad.

Table 2: Conteo de dominios seleccionados

Flag	n
Excluir	59
Incluir	379

FGV para la GEIH de Colombia

Para este proceso se realiza la transformación $\log(\hat{\sigma}_d^2)$ y la selección de las columnas identificador del municipio (`dam2`), la estimación directa (`pobreza`), el número de personas en el dominio (`nd`) y la varianza estimada (`vardir`).

Table 3: Set datos para la FGV

dam2	pobreza	nd	vardir	ln_sigma2
05001	0.1597	27432	0.0000	-10.012
05002	0.4049	257	0.0032	-5.737
05031	0.3817	199	0.0042	-5.463
05034	0.4731	223	0.0018	-6.335
05045	0.2876	480	0.0064	-5.045

Analisis gráfico

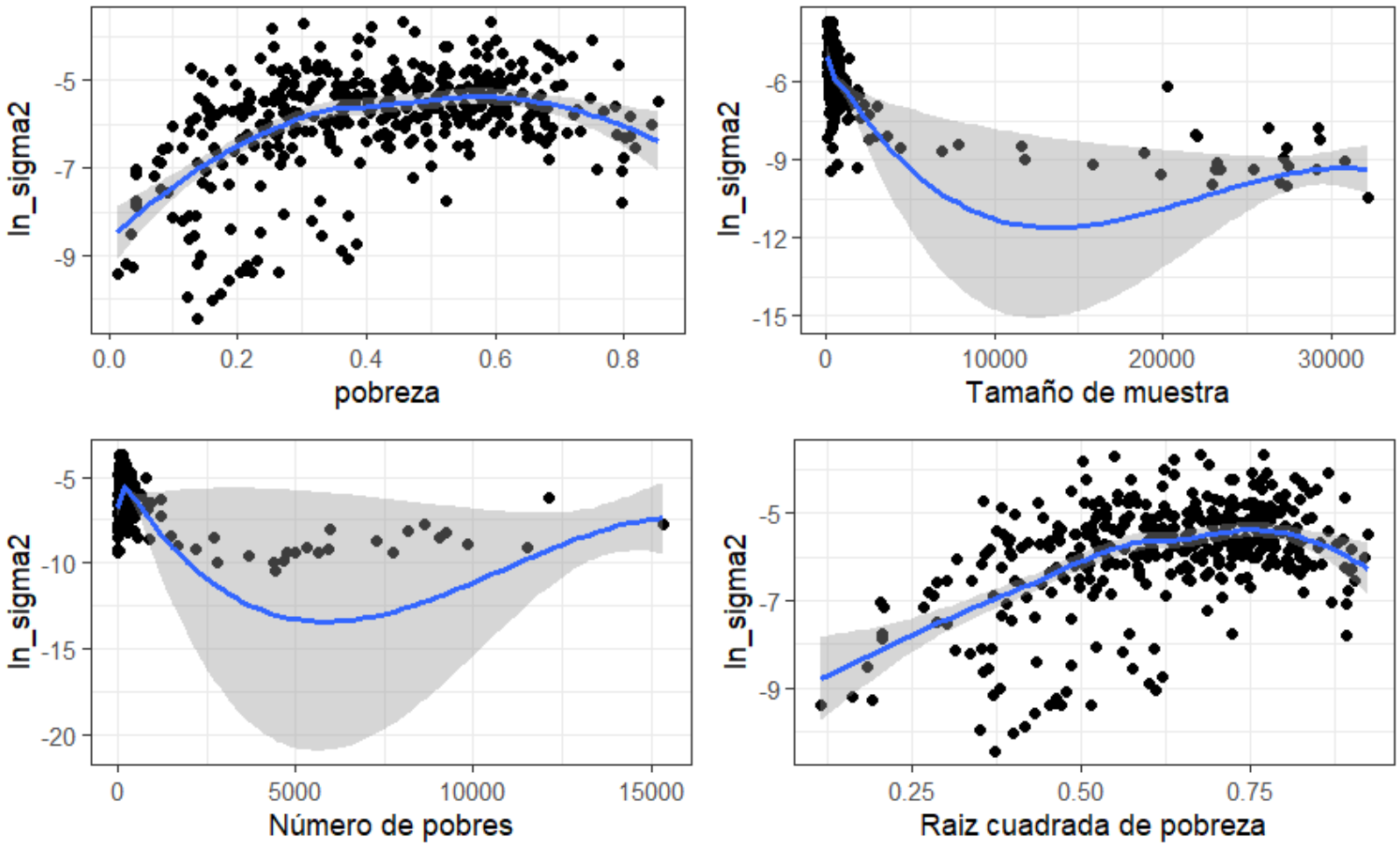

Figure 4: Diagramas de dispersión

Modelo para la varianza

El modelo definido para el conjunto de datos es el siguiente.

$$\log(\hat{\sigma}^2) = \hat{\theta}_{dir} + n_d^2 + \sqrt{\hat{\theta}_{dir}}$$

El resultado del modelo se muestra a continuación:

Table 4: Resumen del modelo

Characteristic	**Beta**	**95% CI**	**p-value**
pobreza	-12	-14, -9.5	<0.001
I(nd^2)	0.00	0.00, 0.00	<0.001
I(sqrt(pobreza))	16	14, 19	<0.001
R^2	0.608		
Adjusted R^2	0.604		

Estimación para Δ y predicción.

Apartir de la estimación del modelo se debe obtener el valor de la constante Δ para lo cual se usa el siguiente código.

```
delta.hat = sum(baseFGV$vardir) /
  sum(exp(fitted.values(FGV1)))
```

Por último se tiene la varianza suavizada.

```
hat.sigma <-
  data.frame(
    dam2 = baseFGV$dam2,
    hat_var = delta.hat * exp(fitted.values(FGV1)))
```

Validación de resultados.

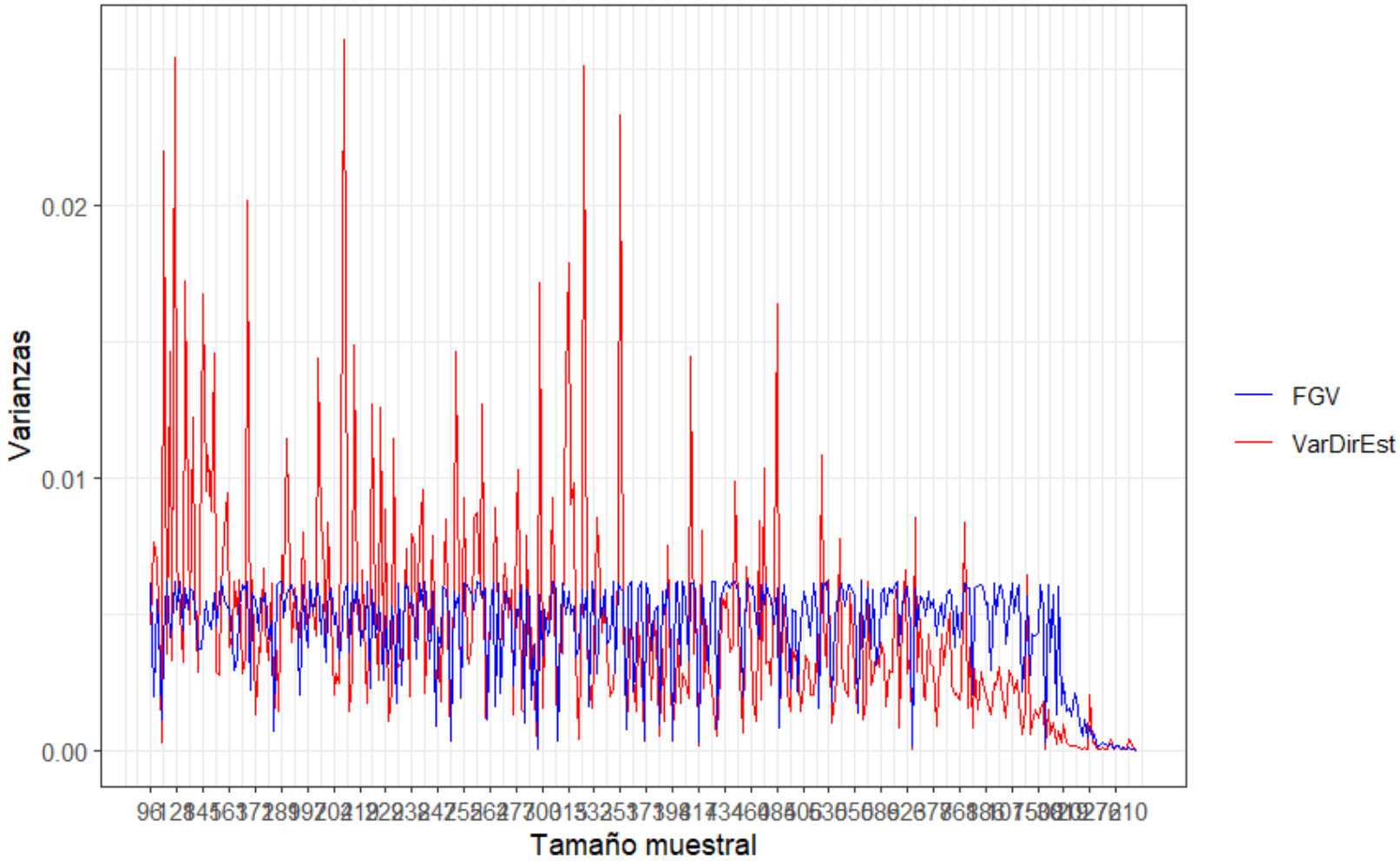

Figure 5: FGV Vs Varianza directa

¡Gracias!

Email: andres.gutierrez@cepal.org