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)

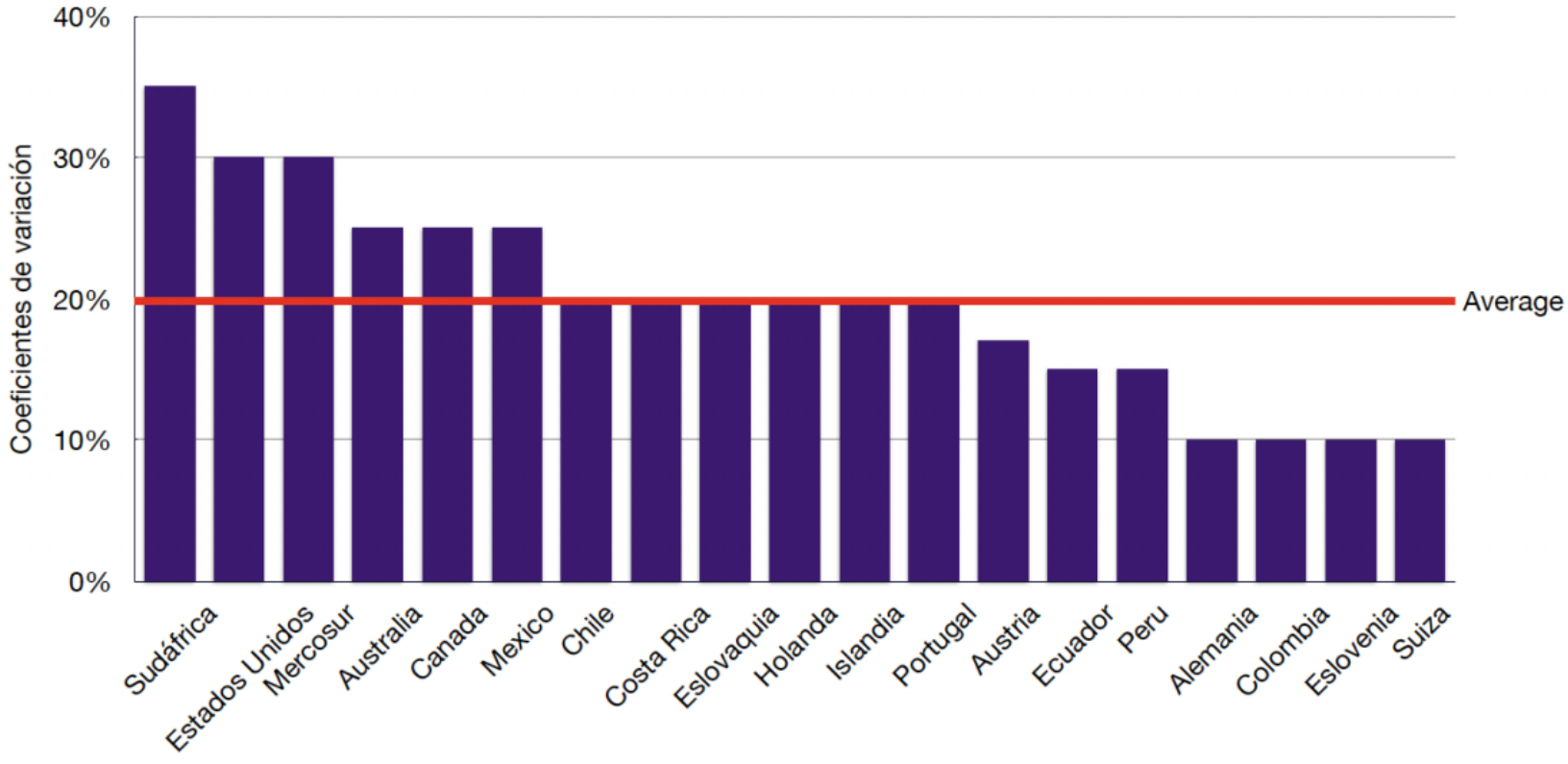


Figure 1: Alertas sobre los coeficientes de variación

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

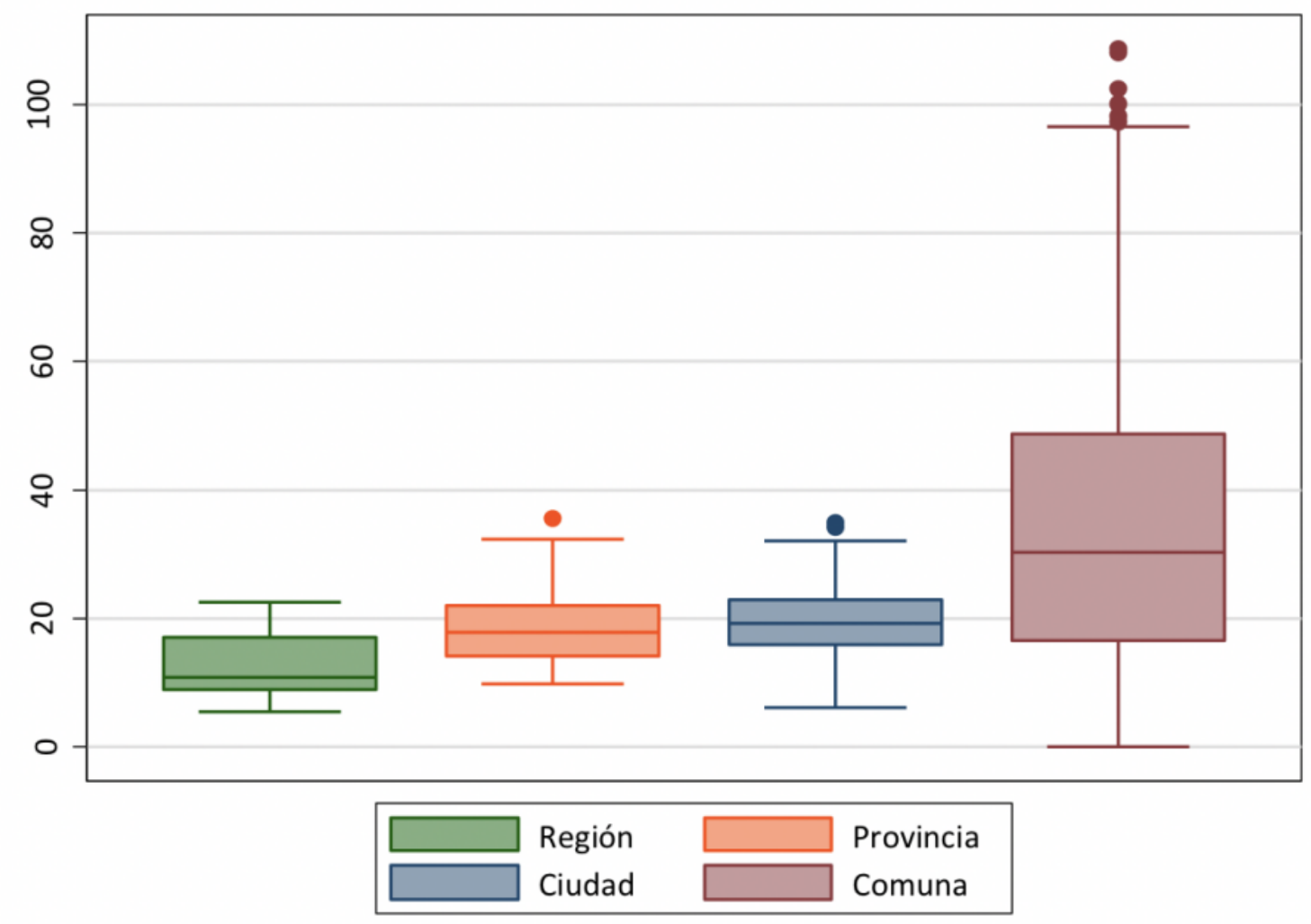


Figure 2: Distribución de los coeficientes de variación en Chile

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

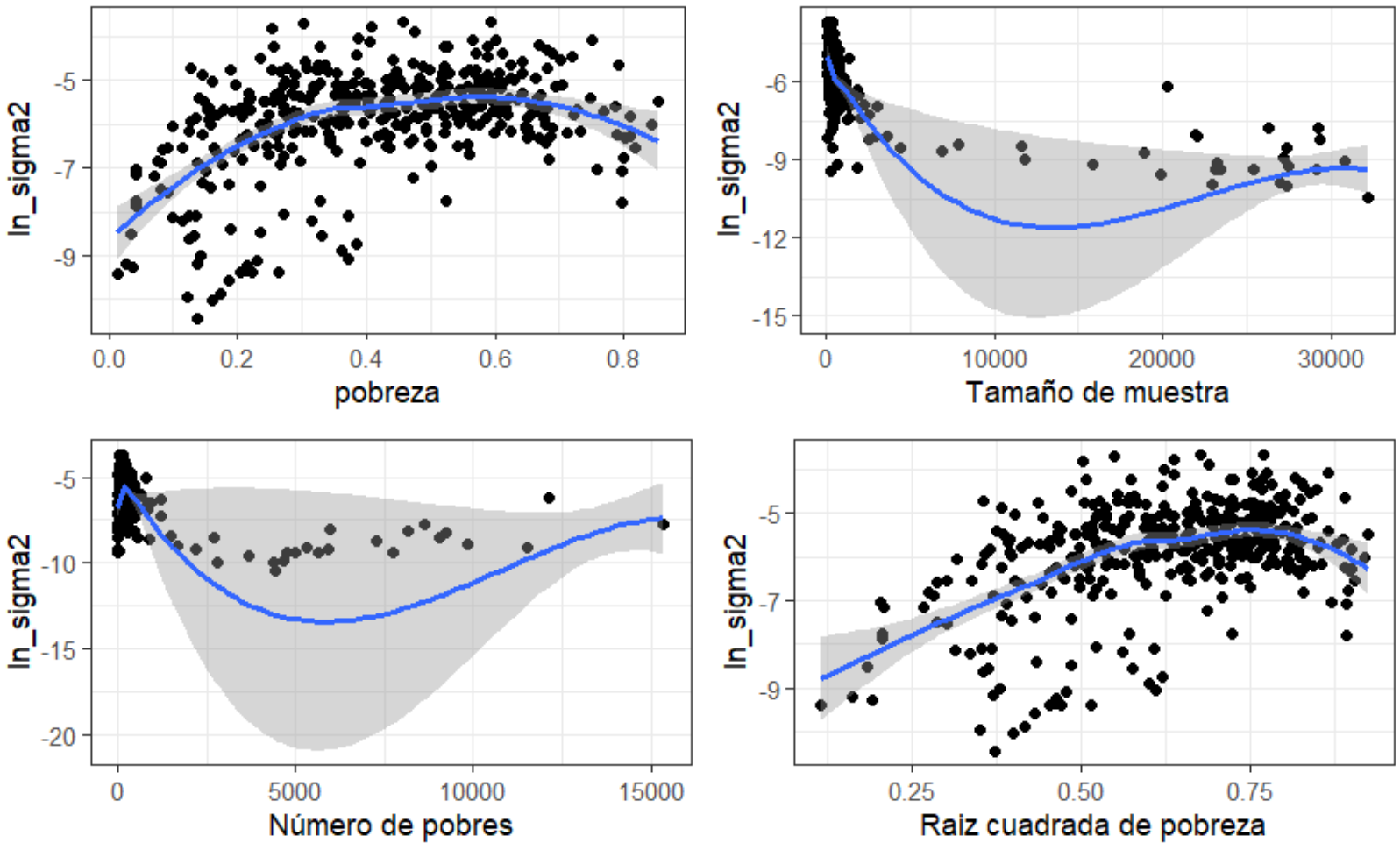


Figure 4: Diagramas de dispersión

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.

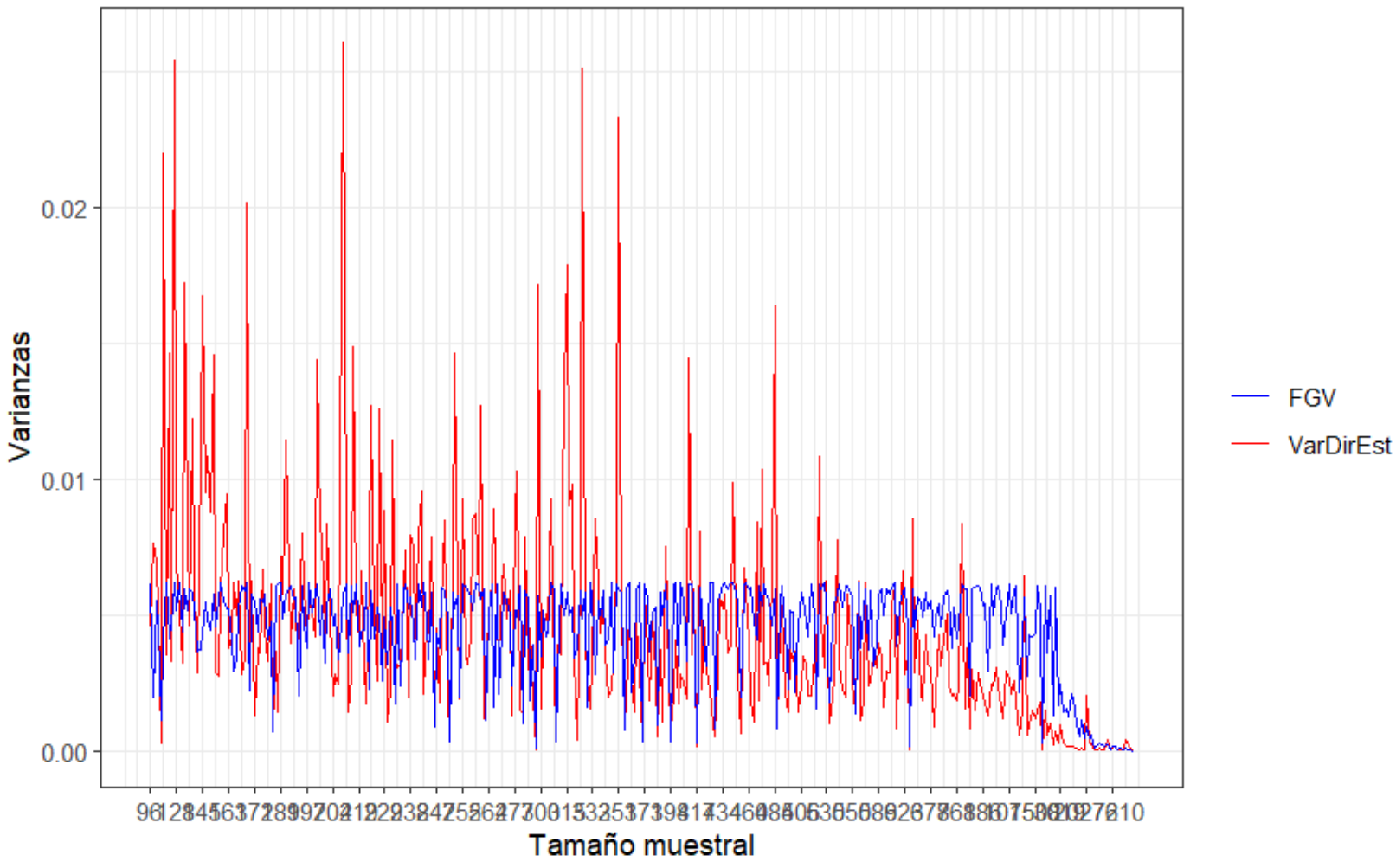


Figure 5: FGV y Varianza directa, por tamaño de muestra

Modelos de área.

Modelo de Fay Herriot

- El Modelo de Fay Herriot, propuesto por Fay y Herriot en 1979, es ampliamente utilizado en la estimación de áreas pequeñas. Este enfoque estadístico se aplica cuando la información a nivel de individuo es limitada, pero se dispone de datos a nivel de áreas y de información auxiliar relacionada con estos datos.
- El modelo establece una relación entre los indicadores de las áreas, θ_d, que varían en función de un vector de covariables x_d. Se formula como $\theta_d = x_d^T \beta + u_d$, donde u_d es un efecto aleatorio específico para cada área.

Modelo de Fay Herriot

- Dado que los valores reales de los indicadores θ_d no son observables, se utiliza el estimador directo $\hat{\theta}_d^{DIR}$ para estimarlos, lo que introduce un error de muestreo. Es decir,

$$\hat{\theta}_d^{DIR} = \theta + e_d$$

- El modelo se ajusta teniendo en cuenta el error de muestreo e_d, y las varianzas $\sigma^2_{e_d}$ se estiman a partir de los microdatos de la encuesta. Esto es:

$$\hat{\theta}_d^{DIR} = x_d^T \beta + u_d + e_d$$

.

Modelo de Fay Herriot

El mejor predictor lineal insesgado (BLUP) bajo el modelo Fay Herriot se calcula como $\tilde{\theta}_d^{FH}$, y se basa en el uso de γ_d para ponderar adecuadamente el estimador directo y la información auxiliar, permitiendo una estimación más precisa de los indicadores en áreas pequeñas. Su ecuación esta dada por:

$$\tilde{\theta}_d^{FH} = x^T d\tilde{\beta} + \tilde{u}_d$$

,

donde $\tilde{u}_d = \gamma_d \left(\hat{\theta}_d^{DIR} - x_d^T \tilde{\beta}\right)$ y $\gamma_d = \frac{\sigma_u^2}{\sigma_u^2 + \sigma_{e_d}^2}$.

Modelo de área para la estimación de la pobreza

Sea P_d la probabilidad de encontrar una persona en condición de pobreza en el $d-$ésimo dominio de la población. Entonces, el estimador directo de P_d se puede escribir como:

$$\hat{P}_d^{DIR} = P_d + e_d$$

Ahora bien, P_d se puede modelar de la siguiente manera,

$$P_d = x_d^T \beta + u_d$$

Modelo de área para la estimación de la pobreza

Reescribiendo $\hat{P}_d^{DIR}$ en términos de las dos ecuaciones anteriores tenemos:

$$\hat{P}_d^{DIR} = x_d^T \beta + u_d + e_d$$

Ahora, es posible suponer que:

- $\hat{P}_d^{DIR} \sim N(x_d^T \beta, \sigma_u^2 + \sigma_{e_d}^2)$,
- $\hat{P}_d^{DIR} \mid u_d \sim N(x_d^T \beta + u_d, \sigma_{e_d}^2)$ y
- $u_d \sim N(0, \sigma_u^2)$

Distribuciones previas.

Las distribuciones previas para β y σ_u^2

$$\beta_p \sim N(0, 10000)$$

$$\sigma_u^2 \sim IG(0.0001, 0.0001)$$

por tanto, el estimador bayesiano para P_d esta dado como $\tilde{P}_d = E\left(P_d \mid \hat{P}_d^{DIR}\right)$

Procedimiento de estimación de la pobreza en los municipios de colombia

Las covariables disponibles se muestran en la siguiente tabla, estas fueron obtenidas previamente.

Table 5: Covariables disponibles

dam	dam2	area1	sexo2	edad2	edad3	edad4	edad5
05	05001	0.9832	0.5299	0.2671	0.2201	0.2355	0.1060
05	05002	0.3953	0.4807	0.2229	0.1977	0.2497	0.1281
05	05004	0.3279	0.4576	0.2376	0.2075	0.2316	0.1218
05	05021	0.5770	0.5020	0.2191	0.1946	0.2357	0.1274
05	05030	0.4859	0.5063	0.2571	0.2047	0.2507	0.0997

Modelo de FH: Rutina en STAN

```
data {
  int<lower=0> N1; // number of data items
  int<lower=0> N2; // number of data items for prediction
  int<lower=0> p;  // number of predictors
  matrix[N1, p] X; // predictor matrix
  matrix[N2, p] Xs; // predictor matrix
  vector[N1] y;    // predictor matrix
  vector[N1] sigma_e; // known variances
}

parameters {
  vector[p] beta;       // coefficients for predictors
  real<lower=0> sigma2_u;
  vector[N1] u;
}
```

Modelo de FH: Rutina en STAN

```
transformed parameters{
  vector[N1] theta;
  vector[N1] thetaSyn;
  vector[N1] thetaFH;
  vector[N1] gammaj;
  real<lower=0> sigma_u;
  thetaSyn = X * beta;
  theta = thetaSyn + u;
  sigma_u = sqrt(sigma2_u);
  gammaj =  to_vector(sigma_u ./ (sigma_u + sigma_e));
  thetaFH = (gammaj) .* y + (1-gammaj).*thetaSyn;
}
```

Modelo de FH: Rutina en STAN

```
model {
  // likelihood
  y ~ normal(theta, sigma_e);
  // priors
  beta ~ normal(0, 100);
  u ~ normal(0, sigma_u);
  sigma2_u ~ inv_gamma(0.0001, 0.0001);
}

generated quantities{
  vector[N2] y_pred;
  for(j in 1:N2) {
    y_pred[j] = normal_rng(Xs[j] * beta, sigma_u);
  }
}
```

Preparando los insumos para STAN

- Definir el modelo de área

```
formula_mod  <- formula(
  ~ sexo2 + anoest2 + anoest3 +
    anoest4 + edad2 + edad3  +  edad4  + edad5 + etnia1 +
    etnia2 + tasa_desocupacion + luces_nocturnas +
    cubrimiento_cultivo + alfabeta
)
```

Preparando los insumos para STAN

- Dividir la base de datos en dominios observados y no observados.

```
# Dominios observados.
data_dir <- base_FH %>% filter(!is.na(pobreza))

Xdat <- model.matrix(formula_mod, data = data_dir)

# Dominios NO observados.
data_syn <-
  base_FH %>% anti_join(data_dir %>% select(dam2))

Xs <- model.matrix(formula_mod, data = data_syn)
```

Preparando los insumos para STAN

- Creando lista de parámetros para STAN

```
sample_data <- list(
  N1 = nrow(Xdat),   # Observados.
  N2 = nrow(Xs),   # NO Observados.
  p  = ncol(Xdat),        # Número de regresores.
  X  = as.matrix(Xdat),  # Covariables Observados.
  Xs = as.matrix(Xs),    # Covariables NO Observados
  y  = as.numeric(data_dir$pobreza), # Estimación directa
  sigma_e = sqrt(data_dir$hat_var)   # Error de estimación
)
```

Compilando el modelo en STAN

La forma de compilar el código de STAN desde R.

```
library(rstan)
fit_FH_normal <- "www/02_FH_Nornal/17FH_normal.stan"
options(mc.cores = parallel::detectCores())
model_FH_normal <- stan(
  file = fit_FH_normal,
  data = sample_data,
  verbose = FALSE,
  warmup = 500,
  iter = 1000,
  cores = 4
)
saveRDS(object = model_FH_normal,
        file = "www/02_FH_Nornal/model_FH_normal.rds")
```

Resultados del modelo para los dominios observados.

Empleando la función `ppc_dens_overlay()`para graficar una comparación entre la distribución empírica de la variable observada pobreza en los datos y las distribuciones predictivas posteriores simuladas para la misma variable.

```
y_pred_B <- as.array(model_FH_normal,
                     pars = "theta") %>%
  as_draws_matrix()

rowsrandom <- sample(nrow(y_pred_B), 100)

y_pred2 <- y_pred_B[rowsrandom,]

ppc_dens_overlay(y = as.numeric(data_dir$pobreza),
                 y_pred2)
```

Chequeo Predictivo Posterior

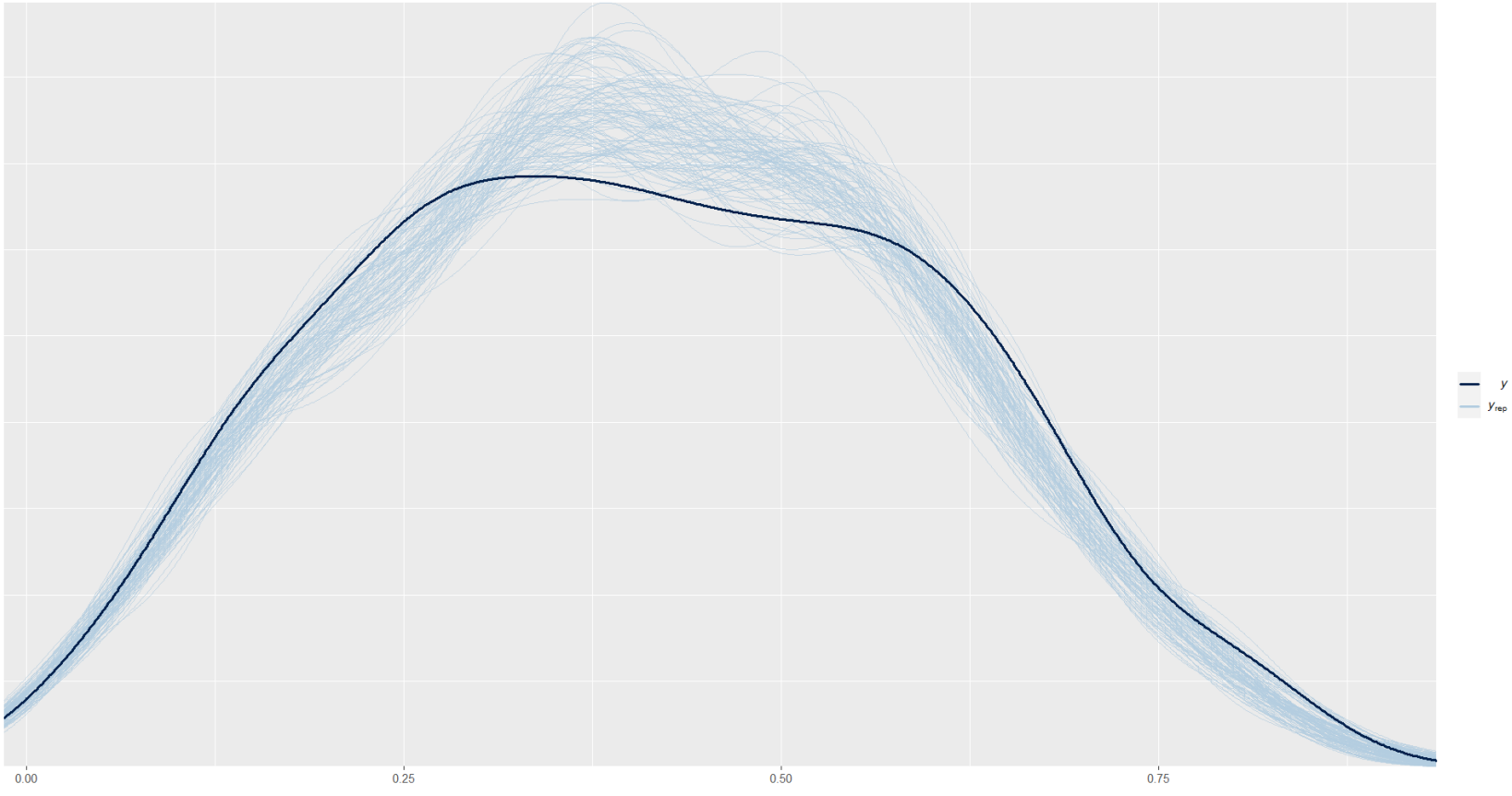


Figure 6: PPC

Validacion de convengencia de cadenas σ^2

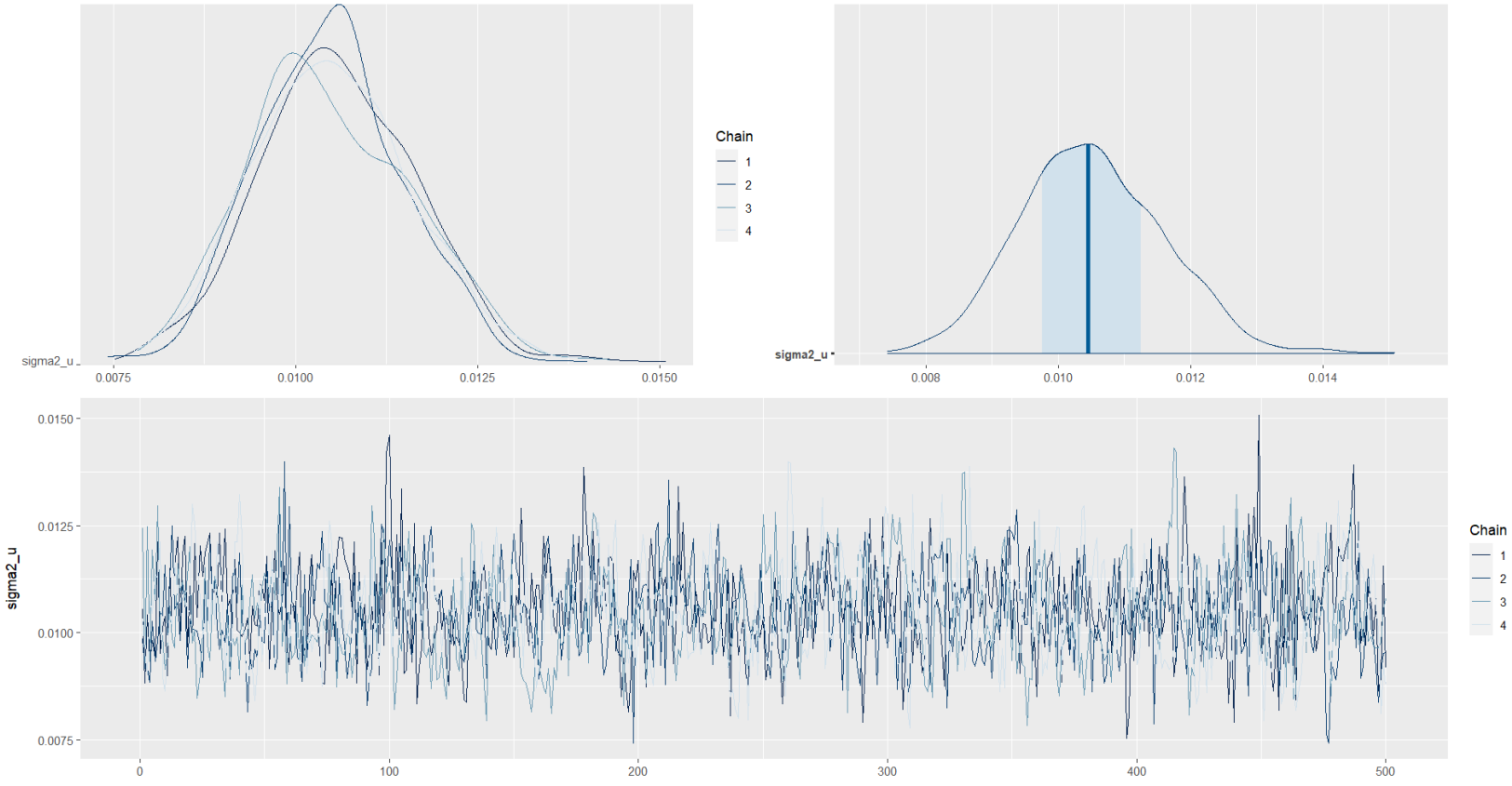
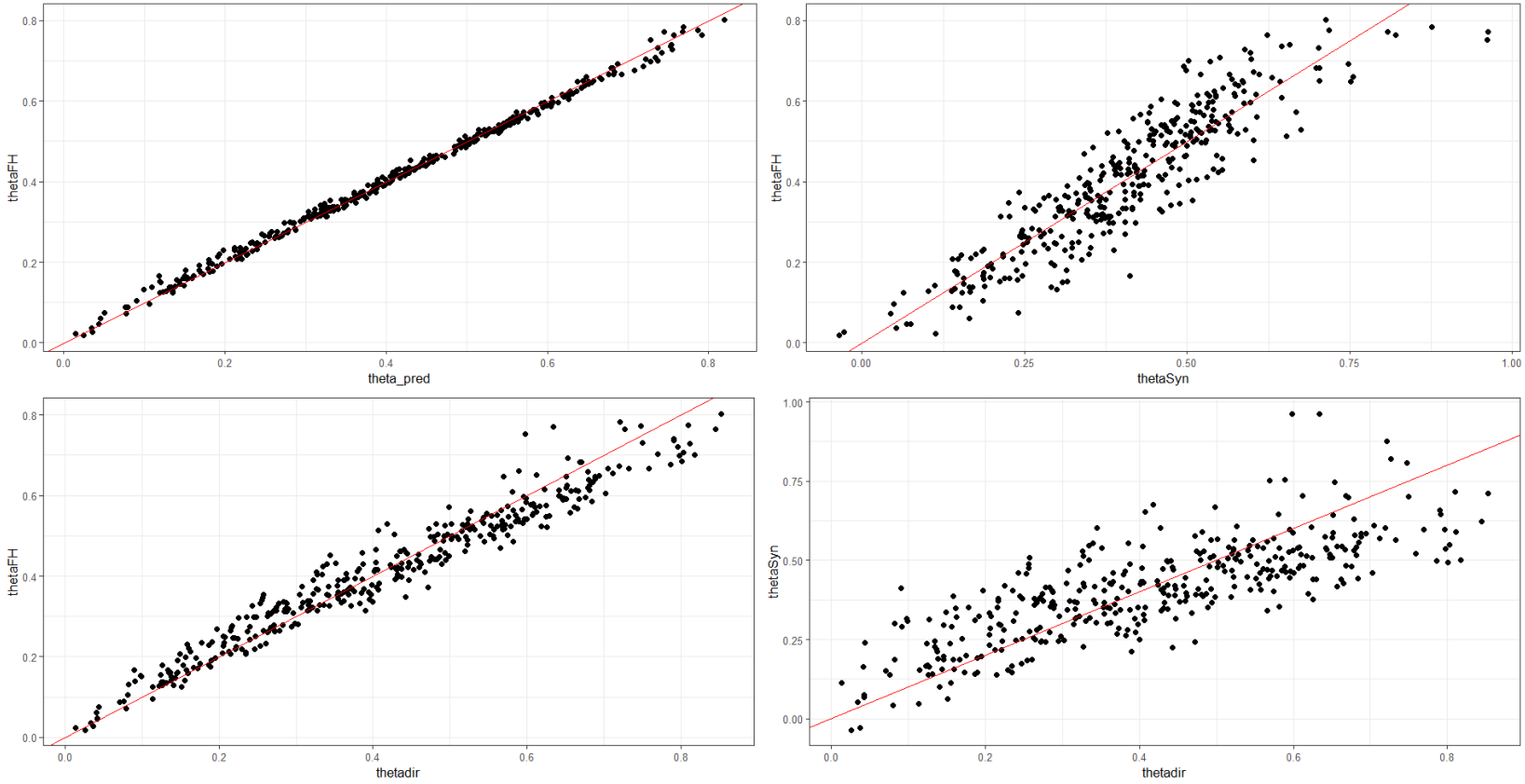


Figure 7: Convergencia de la cadena

Comparación de las estimaciones



Proceso de Benchmarking

- Del censo extraer el total de personas por DAM2

dam	dam2	total_pp	dam_pp
05	05001	2372330	44164417
05	05002	17599	44164417
05	05004	2159	44164417
05	05021	3839	44164417
05	05030	26821	44164417
05	05031	20265	44164417
05	05034	38144	44164417
05	05036	5027	44164417
05	05038	10500	44164417
05	05040	14502	44164417

Estimación directa

Obtener las estimaciones directa por DAM o el nivel de agregación en el cual la encuesta es representativa.

```
directoDam <- diseno %>%
   group_by(Agregado = "Nacional") %>%
  summarise(
    theta_dir = survey_mean(pobreza, vartype = c("ci"))
    )
```

Agregado	theta_dir	theta_dir_low	theta_dir_upp
Nacional	0.2986	0.2935	0.3038

Calculo de ponderadores

Luego de organizar la información anterior se realizaa el calculo de los pesos para el Benchmark

```
estimacionesPre <-
  readRDS("www/02_FH_Nornal/05_tabla_estimacionesPre.rds")
temp <- estimacionesPre %>%
  inner_join(N_dam_pp) %>%
  mutate(theta_dir = directoDam$theta_dir)
R_dam2 <- temp %>%
  summarise(
    R_dam_RB = unique(theta_dir) /
      sum((total_pp  / dam_pp) * theta_pred))
```

R_dam_RB
1.016

Estimación con el modelo de área despues del Benchmarking

```
pesos <- temp %>%
  mutate(W_i = total_pp / dam_pp) %>%
  select(dam2, W_i)

estimacionesBench <- estimacionesPre %>%
  mutate(R_dam_RB = R_dam2$R_dam_RB) %>%
  mutate(theta_pred_RBench = R_dam_RB * theta_pred) %>%
  select(dam, dam2, theta_pred, theta_pred_RBench)
```

dam	dam2	W_i	theta_pred	theta_pred_RBench
05	05001	0.0537	0.1593	0.1618
05	05002	0.0004	0.4130	0.4194
05	05031	0.0005	0.4121	0.4185
05	05034	0.0009	0.4406	0.4474
05	05045	0.0026	0.3118	0.3166

Validación de los resultados.

Este código junta las estimaciones del modelo con pesos de benchmarking con los valores observados y sintéticos, y luego resume las estimaciones combinadas para compararlas con la estimación directa obtenida anteriormente.

```
temp <- estimacionesBench %>%
  left_join(estimacionesPre) %>%
  summarise(
    thetaSyn = sum(W_i * thetaSyn),
    thetaFH = sum(W_i * theta_pred),
    theta_RBench = sum(W_i * theta_pred_RBench)
  ) %>%
  mutate(
    theta_dir = directoDam$theta_dir,
    theta_dir_low = directoDam$theta_dir_low,
    theta_dir_upp = directoDam$theta_dir_upp
  )
```

Resultado de la Validación

Table 6: Comparación de las estimaciones

theta_dir_low	theta_dir_upp	Metodo	Estimacion
0.2935	0.3038	thetaSyn	0.2955
0.2935	0.3038	thetaFH	0.2941
0.2935	0.3038	theta_RBench	0.2986
0.2935	0.3038	theta_dir	0.2986

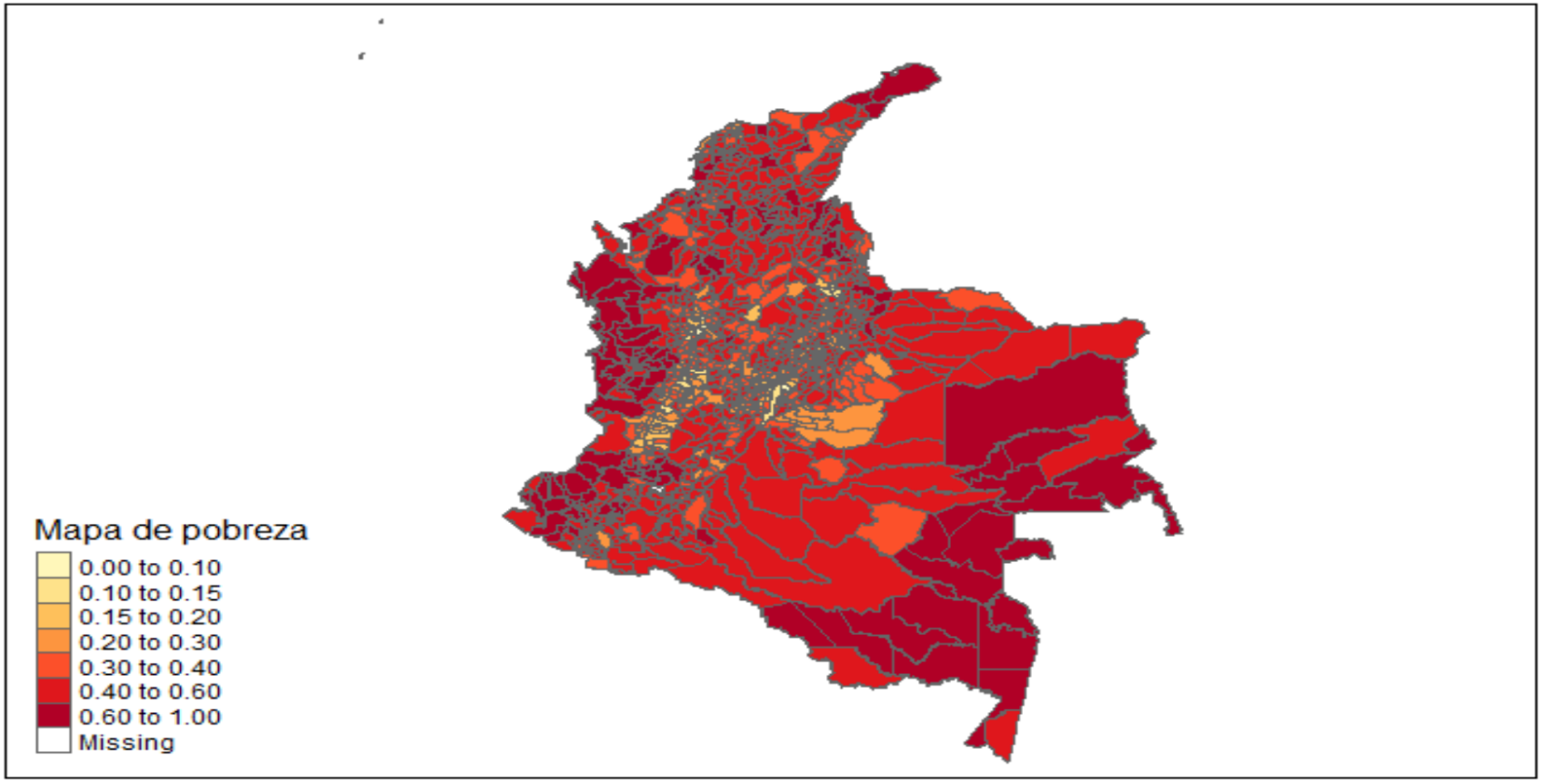


Figure 8: Mapa de pobreza

Modelo de área: Transformación Arcoseno.

- En el modelo de Fay-Herriot, la combinación lineal de covariables puede generar valores que no están dentro del rango aceptable para una proporción.
- Para abordar esto, se aplica una transformación arcoseno a los estimadores: $\hat{z}_d = arcsin\left(\sqrt{\hat{\theta}_d}\right)$.

Varianza de la Transformación Arcoseno

- La varianza de la transformación arcoseno está relacionada con el factor de corrección DEFF y el tamaño de muestra efectivo:

$$Var\left(\hat{z}_d\right) = \frac{\widehat{DEFF}_d}{4 \times n_d} = \frac{1}{4 \times n_{d,efectivo}}$$

.

Especificación del Modelo de Fay-Herriot

- El modelo de Fay-Herriot se define con una variable latente Z_d que sigue una distribución normal.
- La media de Z_d (μ_d) se relaciona con las covariables a través de $x_d^T\beta + u_d$.
- La relación entre la variable latente θ_d y el estimador directo se establece como $\theta_d = (sin(\mu_d))^2$.

Lo anterior se simplifica como:

1. $Z_d \mid \mu_d, \sigma_d^2 \sim N(\mu_d, \sigma_d^2)$
2. $\mu_d = x_d^T\beta + u_d$
3. $\theta_d = (sin(\mu_d))^2$

Distribuciones Previas

Se especifican distribuciones previas para los parámetros del modelo: - $\beta \sim N\left(0, 1000\right)$ - $\sigma_u^2 \sim IG\left(0.0001, 0.0001\right)$.

Modelo de área: Rutina en STAN

El código es similar al anterior, aquí se muestran las variaciones

```
transformed parameters{
  vector[N1] theta;
  vector[N1] lp;
  real<lower=0> sigma_u;
  lp = X * beta + u;
  sigma_u = sqrt(sigma2_u);
  for(k in 1:N1){
    theta[k] = pow(sin(lp[k]), 2);
  }
}
```

Modelo de FH: Rutina en STAN

```
model {
  // likelihood
  y ~ normal(lp, sigma_e);
  // priors
  beta ~ normal(0, 100);
  u ~ normal(0, sigma_u);
  sigma2_u ~ inv_gamma(0.0001, 0.0001);
}
```

Procedimiento de estimación

Para la base preparada previamente hay que seleccionar y transformar las columnas de interés.

```
statelevel_predictors_df <-
  readRDS("www/03_FH_Arcsin/statelevel_predictors.rds")
base_FH <-
  readRDS("www/03_FH_Arcsin/base_FH_2018.rds") %>%
  transmute(
    dam2,          ## id dominios
    pobreza,
    T_pobreza = asin(sqrt(pobreza)),  ## creando zd
    n_effec = n_eff_FGV,        ## n efectivo
    varhat = 1/(4*n_effec)      ## varianza para zd
    )
base_FH <- full_join(base_FH,
            statelevel_predictors_df, by = "dam2" )
```

Preparando los insumos para STAN

Selección de las covariables, que corresponden a las seleccionadas previamente.

```
names_cov <- c(
  "sexo2" , "anoest2" , "anoest3",   "anoest4",
  "edad2" , "edad3" , "edad4" , "edad5" , "etnia1",
  "etnia2" ,  "tasa_desocupacion" , "luces_nocturnas" ,
  "cubrimiento_cultivo" , "alfabeta"
)
```

Dividir el set de datos.

El proceso de estimación y predicción se hace por separado dentro de STAN

- Dominios observados.

```
data_dir <- base_FH %>% filter(!is.na(T_pobreza))
Xdat <- cbind(inter = 1,data_dir[,names_cov])
```

- Dominios NO observados.

```
data_syn <-
  base_FH %>% anti_join(data_dir %>% select(dam2))
Xs <-  cbind(inter = 1,data_syn[,names_cov])
```

Lista de parámetros para STAN

El motor de procesamiento de STAN se basa en C++, por lo que hace necesario que los argumentos para ejecutar los códigos ingresen en forma de lista.

```
sample_data <- list(
  N1 = nrow(Xdat),          # Observados.
  N2 = nrow(Xs),            # NO Observados.
  p  = ncol(Xdat),          # Número de regresores.
  X  = as.matrix(Xdat),     # Covariables Observados.
  Xs = as.matrix(Xs),       # Covariables NO Observados
  y  = as.numeric(data_dir$T_pobreza),
  sigma_e = sqrt(data_dir$varhat)
)
```

Compilando el modelo en STAN

```
fit_FH_arcoseno <-
  "www/03_FH_Arcsin/15FH_arcsin_normal.stan"

model_FH_arcoseno <- stan(
  file = fit_FH_arcoseno,
  data = sample_data,
  verbose = FALSE,
  warmup = 500,
  iter = 1000,
  cores = 4
)
saveRDS(model_FH_arcoseno,
        "www/03_FH_Arcsin/model_FH_arcoseno.rds")
```

Resultados del modelo para los dominios observados.

De forma similar al modelo de Fay Herrior se realiza el gráfico con el chequeo predictivo posterior.

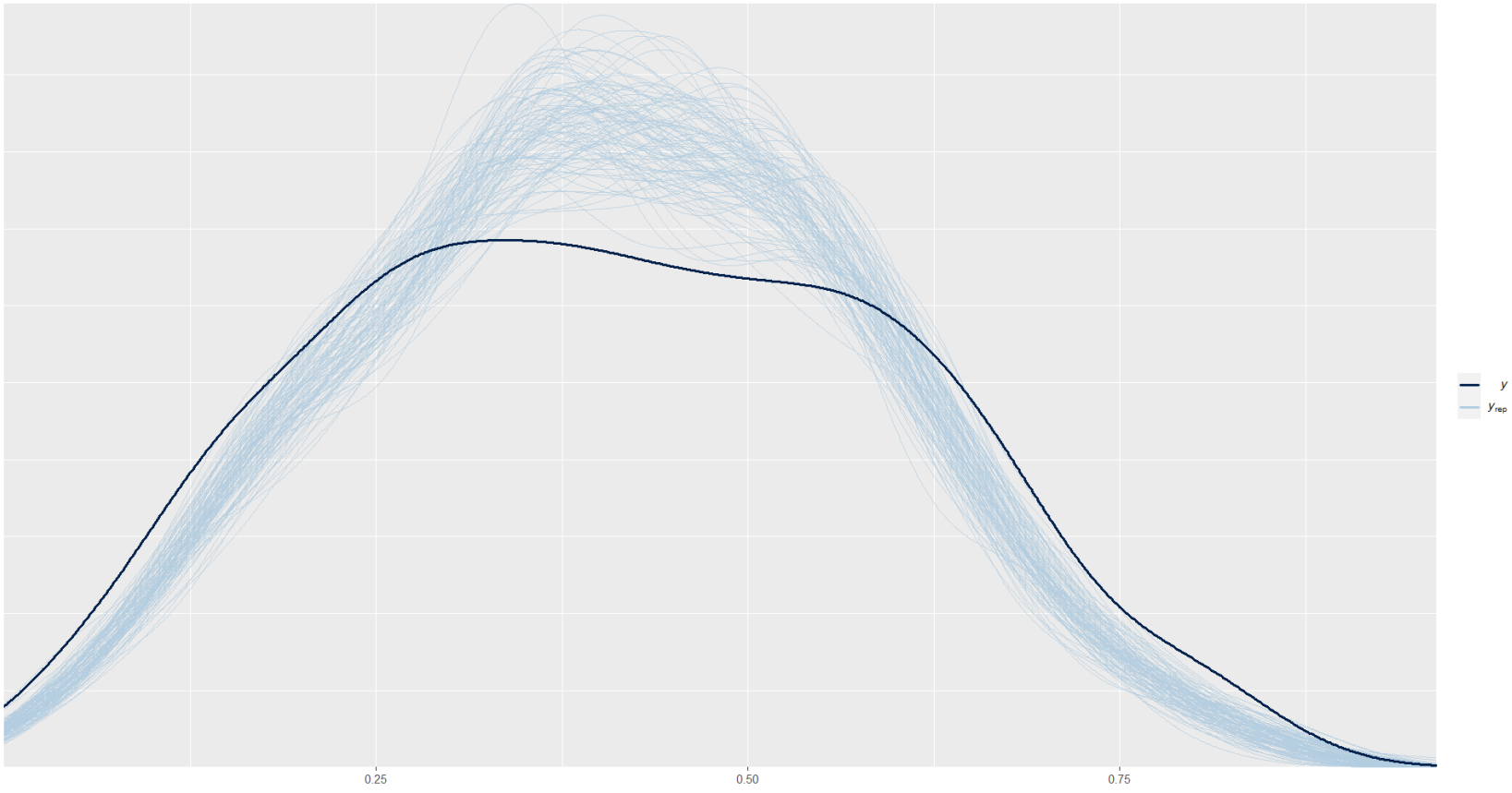


Figure 9: PPC Arcosin

Análisis gráfico de la convergencia de las cadenas de σ_u^2.

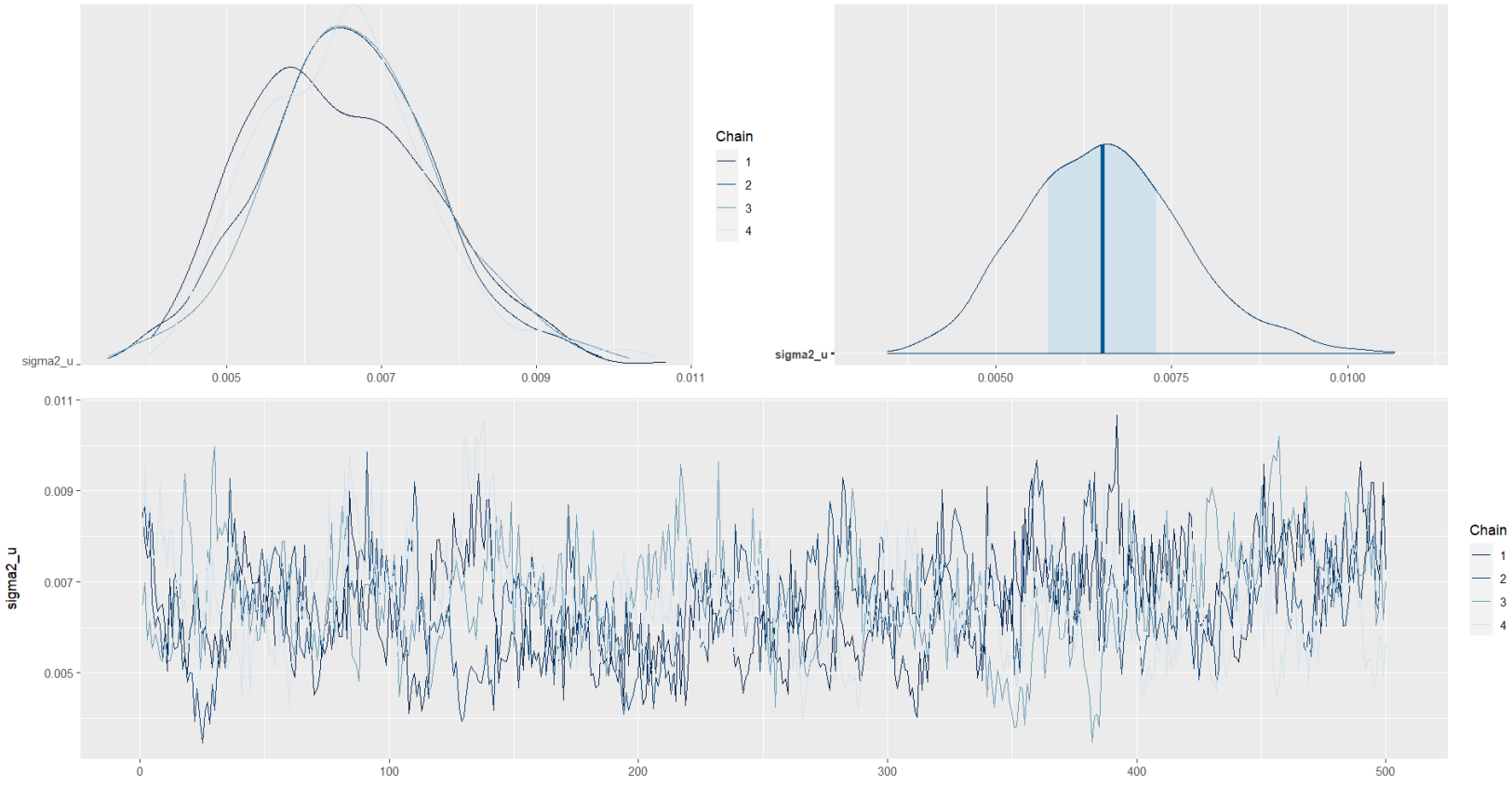


Figure 10: Recorrido de las cadenas

Mapa de pobreza con transformación Arcosin

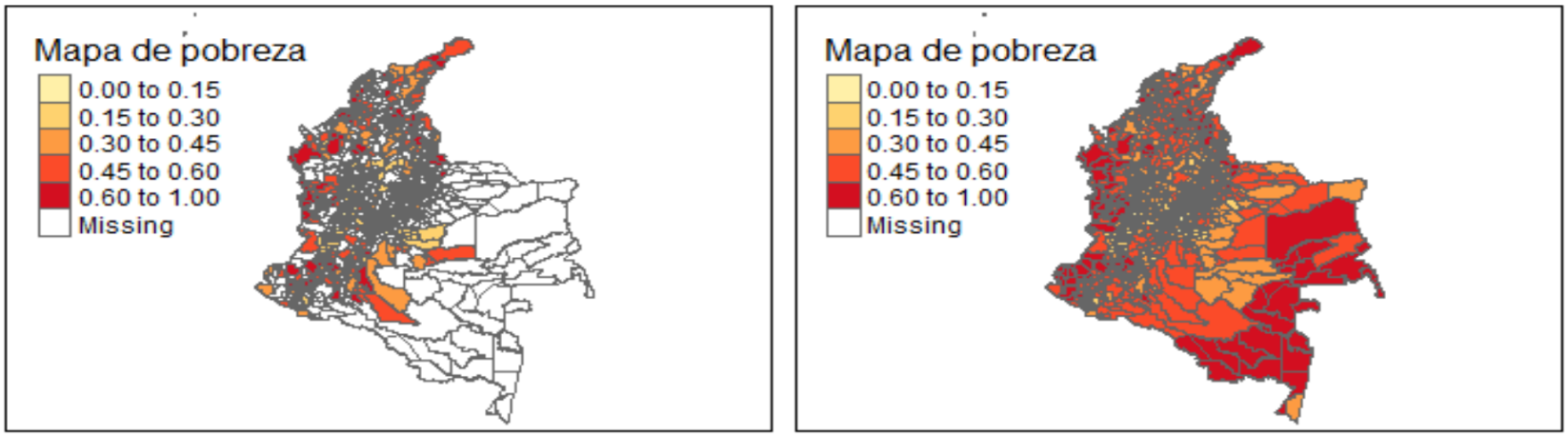


Figure 11: Mapa de pobreza con transformación Arcosin

Mapa de los coeficientes de variación para la pobreza

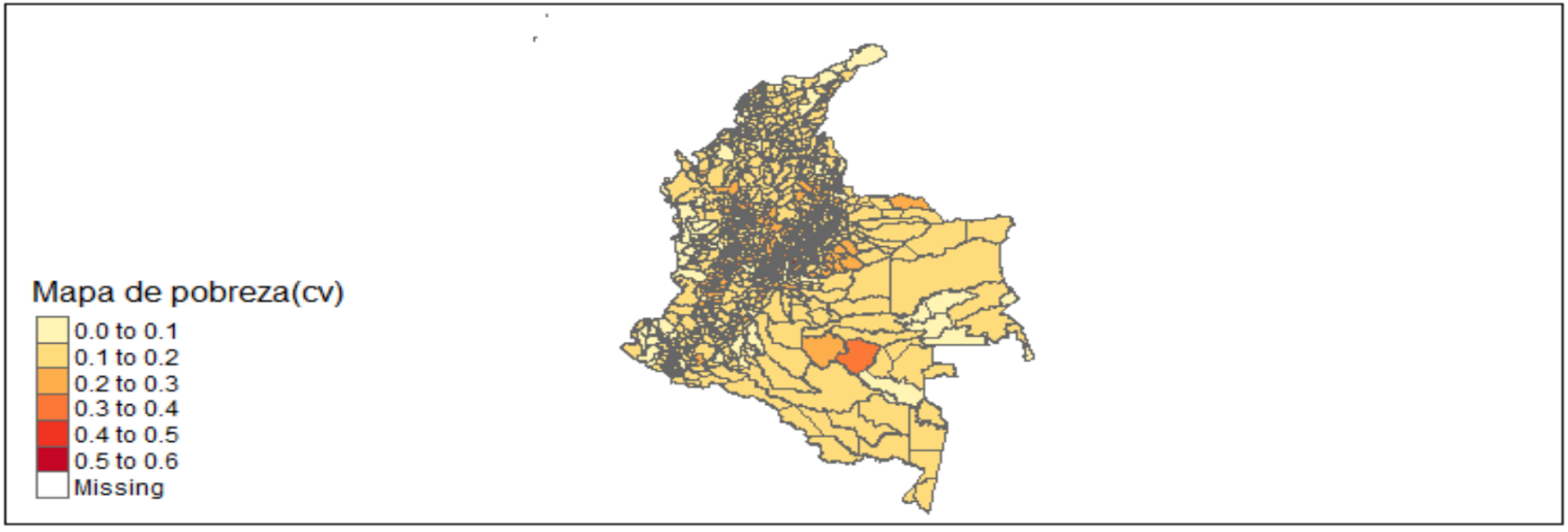


Figure 12: Mapa de los coeficientes de variación

Modelos de área con variable respuesta Beta.

- El modelo beta-logístico se introdujo inicialmente en el contexto de un enfoque de Estimación de Mejor Predicción (EBP) por Jiang y Lahiri en 2006. Fue utilizado para estimar medias de dominio en poblaciones finitas.
- El modelo área beta-logístico se define a través de la siguiente expresión:
  - $\hat{p}_d \mid P_d \sim beta(a_d, b_d)$.
  - La función de enlace se relaciona con los parámetros del modelo:
    - $logit(P_d) \mid \beta, \sigma_u^2 \sim N(x_d^T \beta, \sigma_u^2)$.

Estimación de Parámetros

- Los parámetros a_d y b_d se estiman de la siguiente manera:
  - $a_d = P_d \times \phi_d$
  - $b_d = (1 - P_d) \times \phi_d$
- Donde $\phi_d = \frac{n_d}{\widehat{DEFF}_d} - 1 = n_{d,efectivo} - 1$.
- Se especifican distribuciones previas para los parámetros del modelo:
  - $\beta_k \sim N(0, 10000)$
  - $\sigma_u^2 \sim IG(0.0001, 0.0001)$.

Modelo de área: Rutina en STAN

En este bloque de código vemos la transformación que se realiza sobre los parámetros de entrada.

```
transformed parameters{
  vector[N1] LP;
  real<lower=0> sigma_u;
  vector[N1] theta;
  LP = X * beta + u;
  sigma_u = sqrt(sigma2_u);
  for (i in 1:N1) {
    theta[i] = inv_logit(LP[i]);
  }
}
```

Modelo de FH: Rutina en STAN

```
model {
  // model calculations
  vector[N1] a;
  vector[N1] b;

  for (i in 1:N1) {
    a[i] = theta[i] * phi[i];
    b[i] = (1 - theta[i]) * phi[i];
  }

  // priors
  beta ~ normal(0, 100);
  u ~ normal(0, sigma_u);
  sigma2_u ~ inv_gamma(0.0001, 0.0001);

  // likelihood
  y ~ beta(a, b);
}
```

Procedimiento de estimación

En forma similar a los modelos anteriores hacemos uso de la base previamente preparada

```
base_FH <-
readRDS("www/04_FH_Beta_y_Binomial/base_FH_2018.rds") %>%
  select(dam2, pobreza, n_eff_FGV)

base_FH <- full_join(base_FH,
                     statelevel_predictors_df, by = "dam2")
```

Las covariables son las mismas que se emplearon en los modelos anteriores.

Dividir el set de datos.

El proceso de estimación y predicción se hace por separado dentro de STAN

- Dominios observados.

```
data_dir <- base_FH %>% filter(!is.na(T_pobreza))
Xdat <- cbind(inter = 1,data_dir[,names_cov])
```

- Dominios NO observados.

```
data_syn <-
  base_FH %>% anti_join(data_dir %>% select(dam2))
Xs <-  cbind(inter = 1,data_syn[,names_cov])
```

Lista de parámetros para STAN

```
sample_data <- list(
  N1 = nrow(Xdat),   # Observados.
  N2 = nrow(Xs),   # NO Observados.
  p  = ncol(Xdat),        # Número de regresores.
  X  = as.matrix(Xdat),  # Covariables Observados.
  Xs = as.matrix(Xs),    # Covariables NO Observados
  y  = as.numeric(data_dir$pobreza),
  phi = data_dir$n_eff_FGV - 1
)
```

Compilando el modelo en STAN

```
fit_FH_beta_logitic <-
  "www/04_FH_Beta_y_Binomial/16FH_beta_logitc.stan"

model_FH_beta_logitic <- stan(
  file = fit_FH_beta_logitic,
  data = sample_data,
  verbose = FALSE,
  warmup = 500,
  iter = 1000,
  cores = 4
)
saveRDS(model_FH_beta_logitic,
file = "www/04_FH_Beta_y_Binomial/model_FH_beta.rds")
```

Resultados del modelo para los dominios observados.

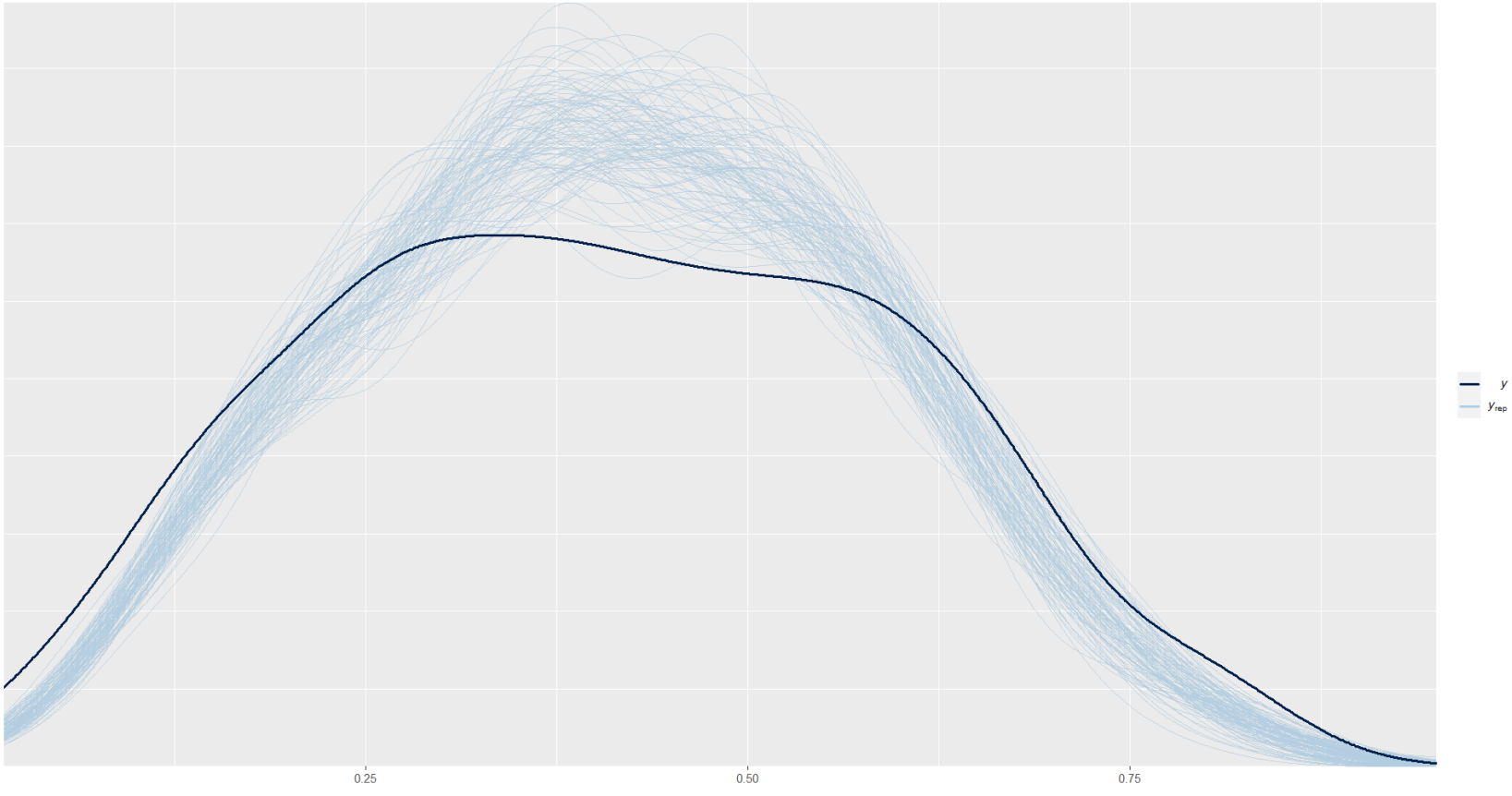


Figure 13: PPC modelo de área Beta

Análisis gráfico de la convergencia de las cadenas de σ_u^2.

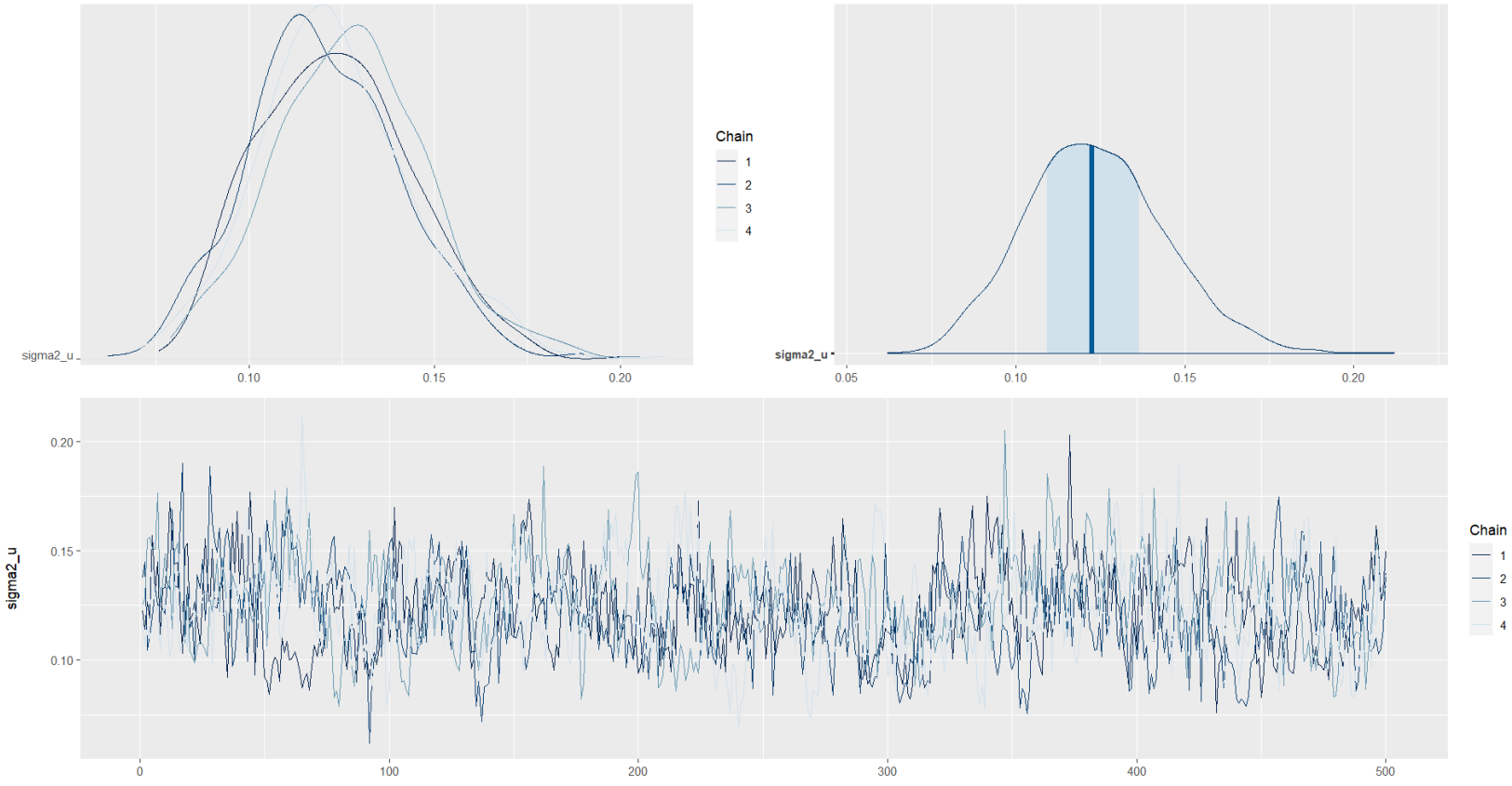


Figure 14: Recorrido de las cadenas

Mapa de pobreza con modelo de área de respuesta beta.

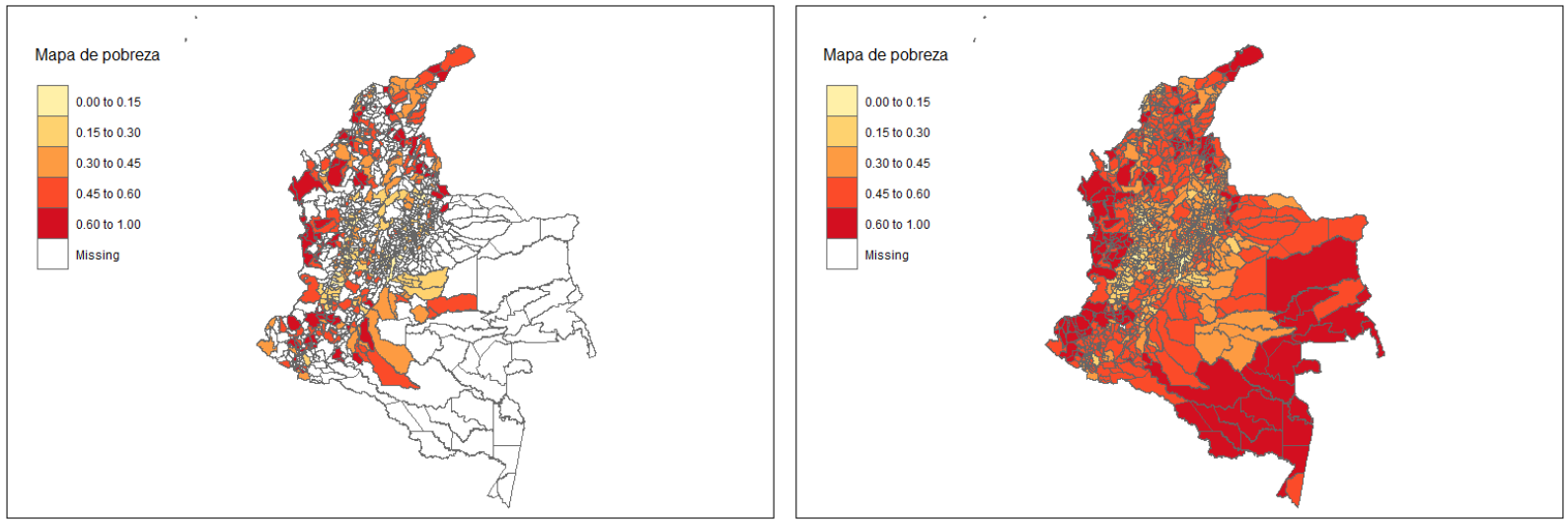


Figure 15: Mapa de pobreza con el modelo de área de respuesta beta.

Mapa de los coeficientes de variación para la pobreza

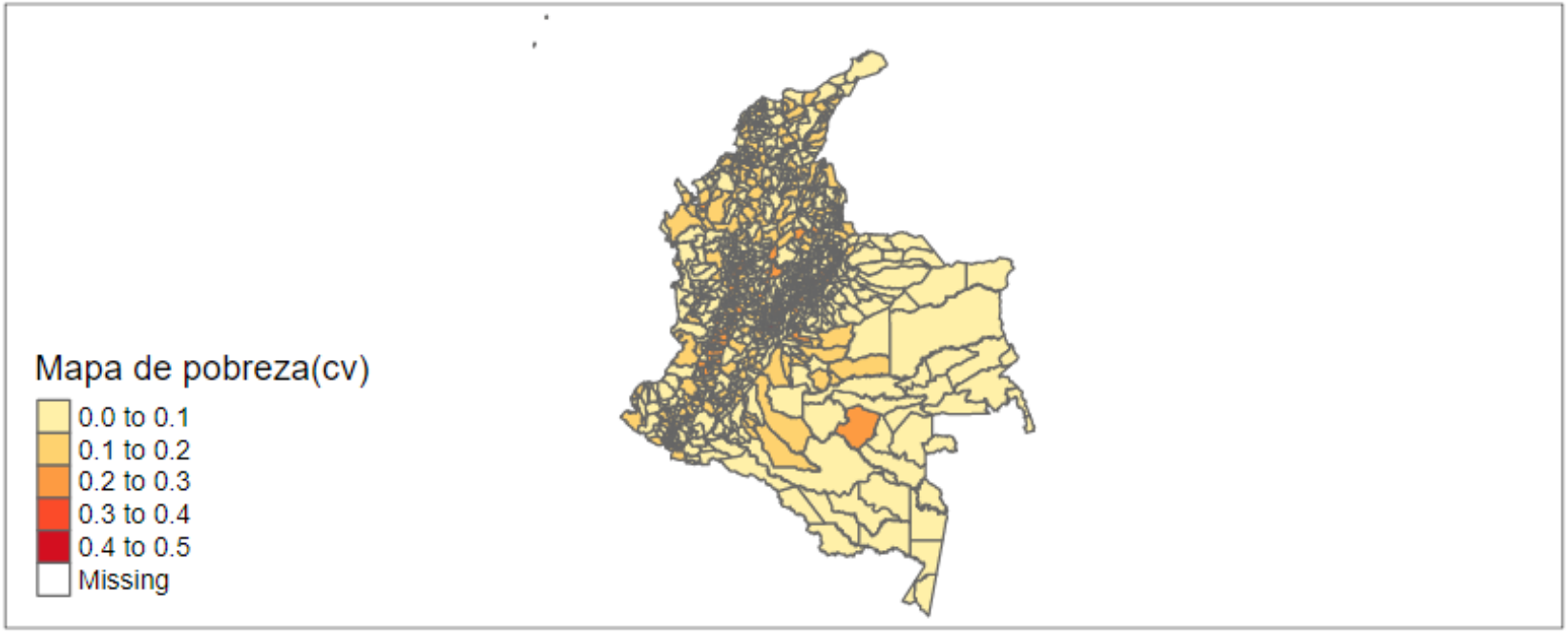


Figure 16: Mapa de los coeficientes de variación

Modelos de área con variable respuesta Binomial.

- El modelo de área de Fay-Herriot puede ser sustituido por un Modelo Mixto Lineal Generalizado (GLMM) cuando los datos observados son inherentemente discretos, como recuentos de personas u hogares con ciertas características.
- En un GLMM, se asume una distribución binomial para los datos Y_d con probabilidad de éxito θ_d y un modelo logístico para θ_d con errores normales en la escala logit.

Ecuación del modelo

El modelo se formula de la siguiente manera:

- $Y_d \mid \theta_d, n_d \sim Binomial(n_d, \theta_d)$
- $logit\left(\theta_d\right) = \log\left(\frac{\theta_d}{1-\theta_d}\right) = x_d^T\beta + u_d$

donde $u_d \sim N\left(0, \sigma_u^2\right)$ y n_d es el tamaño de la muestra para el área d.

Consideraciones para el modelo

Para muestras complejas, surgen dos problemas:

- Los valores de Y_d no son enteros y se ven afectados por las ponderaciones de la encuesta.
- La varianza muestral en la distribución binomial no es precisa.

Propuesta de Carolina Franco.

- Se introduce un **tamaño de muestra efectivo** $\tilde{n}_d$ y un **número de muestra efectivo de éxitos** $\tilde{Y}_d$ para abordar estos problemas y mantener la estimación directa de la pobreza y su varianza correspondiente.
- Dado lo anterior, es posible suponer que

$$\tilde{n}_d \sim \frac{\check{\theta}_d\left(1-\check{\theta}_d\right)}{\widehat{Var}\left(\hat{\theta}_d\right)}$$

  con $\check{\theta}_d$ es una preliminar predicción basada en el modelo para la proporción poblacional, $\hat{\theta}_i$ la estimación directa y $\widehat{Var}(\hat{\theta}_d)$ la estimación de la varianza de muestreo.
- Luego, se asume que $\tilde{n}_d$ es proporcional a la varianza ajustada y que $\tilde{Y}_d = \tilde{n}_d \times \hat{\theta}_d$.

Distribuciones previas

- Se especifican las distribuciones previas para los parámetros β y σ_u^2:
  - $\beta \sim N(0, 10000)$
  - $\sigma_u^2 \sim IG(0.0001, 0.0001)$

Modelo de área: Rutina en STAN

En este bloque de código vemos la transformación que se realiza sobre los parámetros de entrada.

```
transformed parameters {
   vector[N1] LP;
   vector[N1] theta;
   real<lower=0> sigma_u;

   sigma_u = sqrt(sigma2_u);
   LP =  X * beta + u;
   theta = inv_logit(LP);
}
```

Modelo de FH: Rutina en STAN

```
model {
  to_vector(beta) ~ normal(0, 10000);
   u ~ normal(0, sigma_u);
  sigma2_u ~ cauchy(0, 1000);
  for(ii in 1:N1){
  y_effect[ii] ~ binomial(n_effec[ii], theta[ii]);
}
  }
```

Procedimiento de estimación

Lectura de la base de datos con las estimaciones directas.

```
base_FH <-
readRDS("www/04_FH_Beta_y_Binomial/base_FH_2018.rds") %>%
  select(dam2, pobreza, n_eff_FGV)

base_FH <- full_join(base_FH,
                     statelevel_predictors_df, by = "dam2")
```

Las covariables son las mismas que se emplearon en los modelos anteriores.

Dividir el set de datos.

El proceso de estimación y predicción se hace por separado dentro de STAN

- Dominios observados.

```
data_dir <- base_FH %>% filter(!is.na(T_pobreza))
Xdat <- cbind(inter = 1,data_dir[,names_cov])
```

- Dominios NO observados.

```
data_syn <-
  base_FH %>% anti_join(data_dir %>% select(dam2))
Xs <-  cbind(inter = 1,data_syn[,names_cov])
```

Obteniendo Parámetros adicionales.

- Tamaño de muestra efectivo $\tilde{n}_d$

```
n_effec = round(data_dir$n_eff_FGV)
```

- Número de muestra efectivo de éxitos $\tilde{Y}_d$

```
y_effect  = round((data_dir$pobreza)*n_effec)
```

Lista de parámetros para STAN

```
sample_data <- list(
  N1 = nrow(Xdat),   # Observados.
  N2 = nrow(Xs),   # NO Observados.
  p  = ncol(Xdat),        # Número de regresores.
  X  = as.matrix(Xdat),  # Covariables Observados.
  Xs = as.matrix(Xs),    # Covariables NO Observados
  n_effec = n_effec,
  y_effect  = y_effect  # Estimación directa.
)
```

Compilando el modelo en STAN

```
fit_FH_binomial <-
  "www/04_FH_Beta_y_Binomial/14FH_binomial.stan"

model_FH_Binomial <- stan(
  file = fit_FH_binomial,
  data = sample_data,
  verbose = FALSE,
  warmup = 500,
  iter = 1000,
  cores = 4
)

saveRDS(model_FH_Binomial,
file = "www/04_FH_Beta_y_Binomial/model_FH_Binomial.rds")
```

Resultados del modelo para los dominios observados.

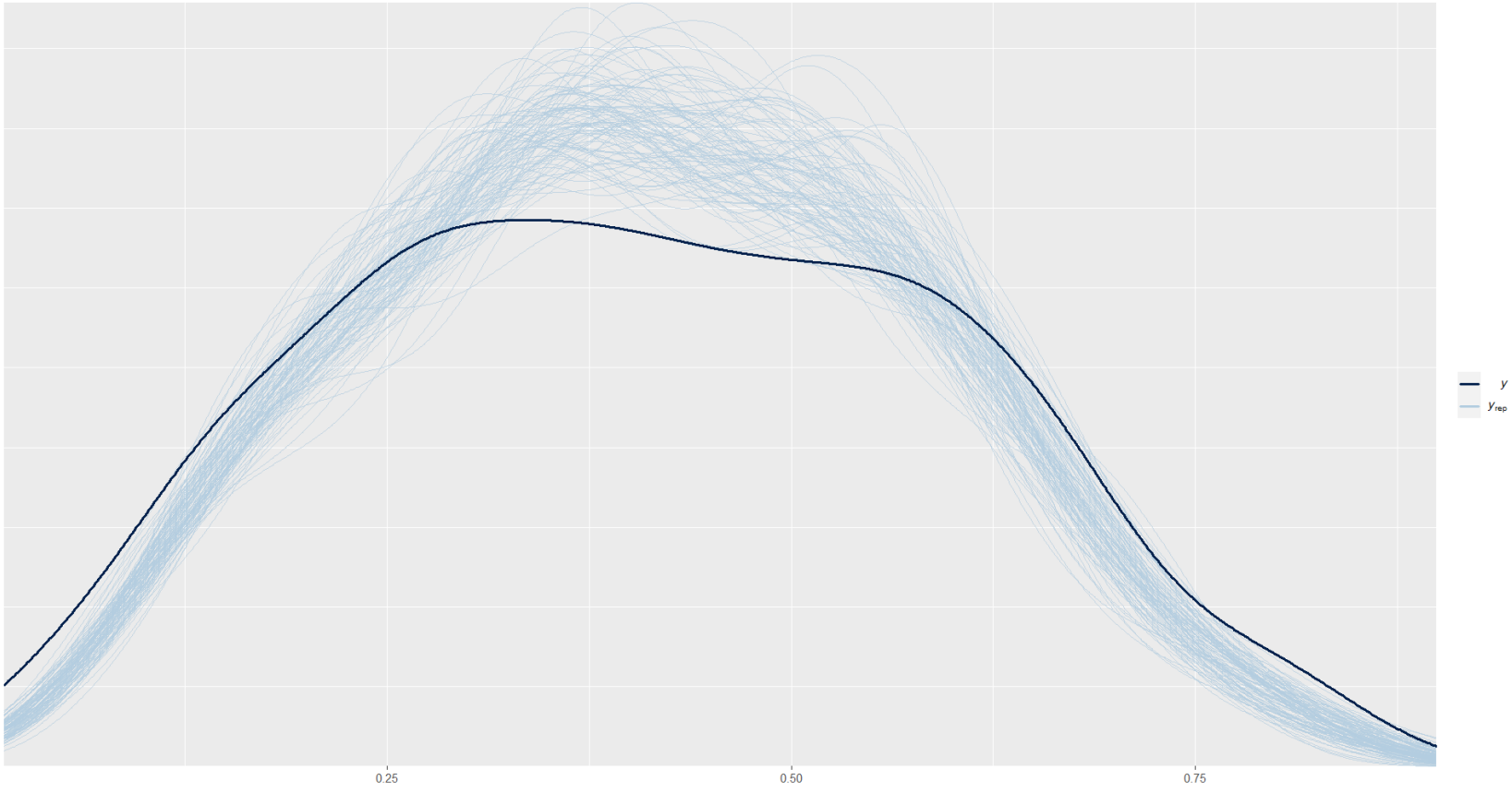


Figure 17: PPC modelo de área Binomial

Análisis gráfico de la convergencia de las cadenas de σ_u^2.

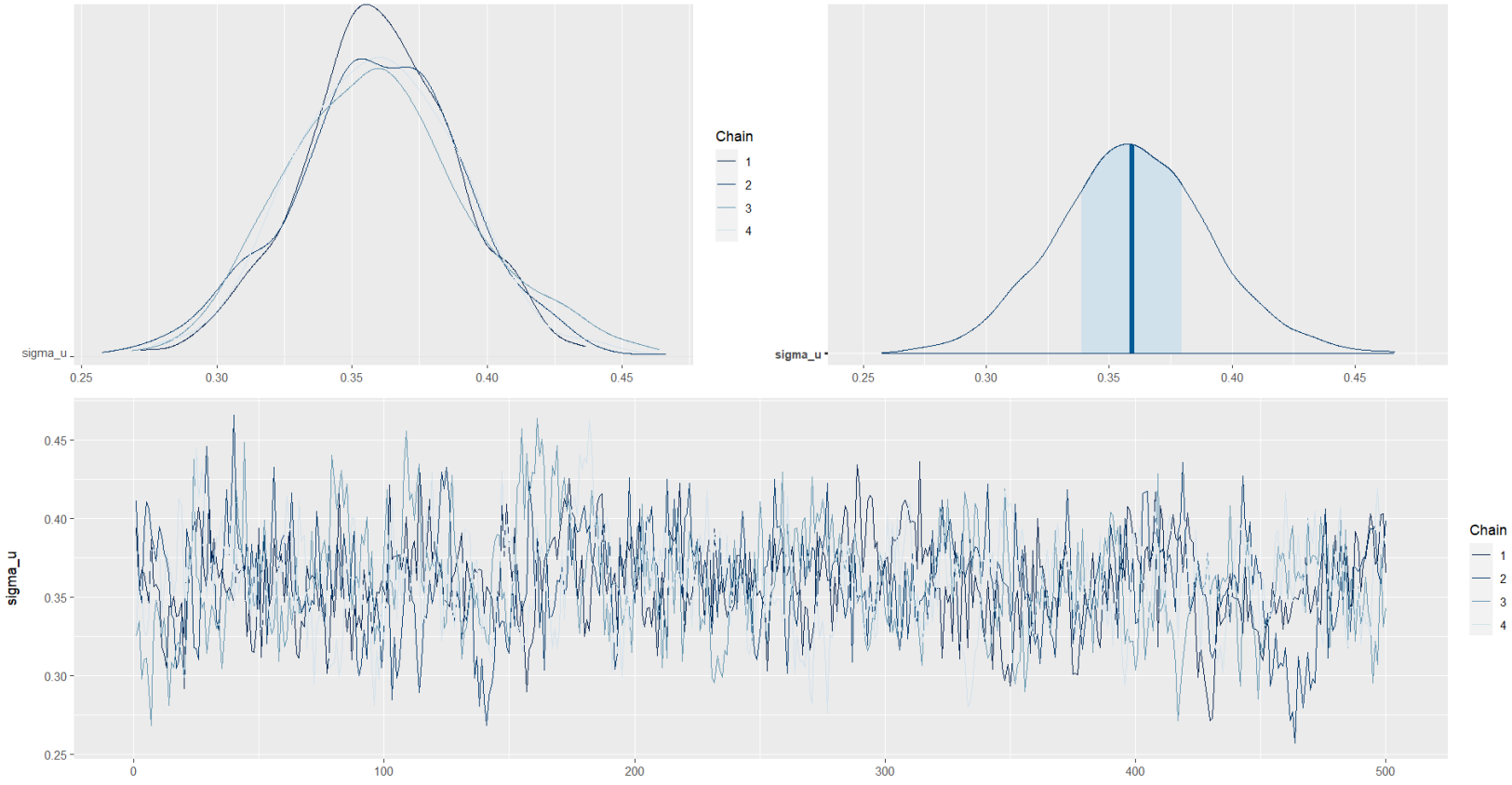


Figure 18: Recorrido de las cadenas

Mapa de pobreza con modelo de área de respuesta binomial

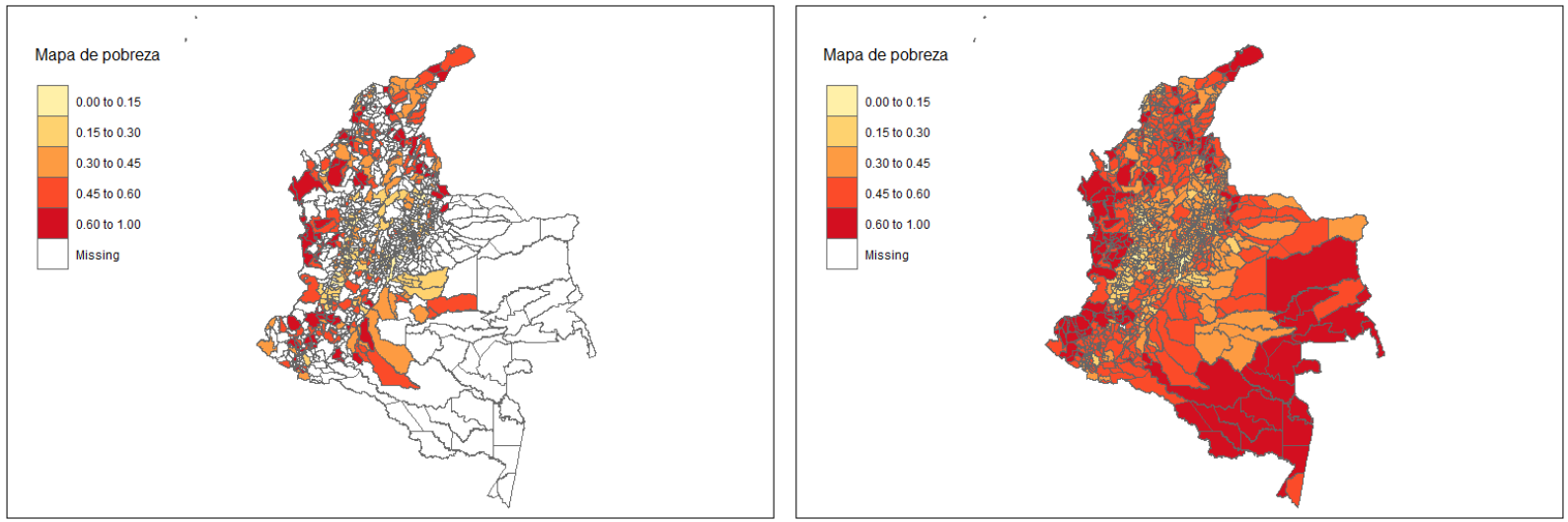


Figure 19: Mapa de pobreza con el modelo de área de respuesta beta.

Mapa de los coeficientes de variación para la pobreza

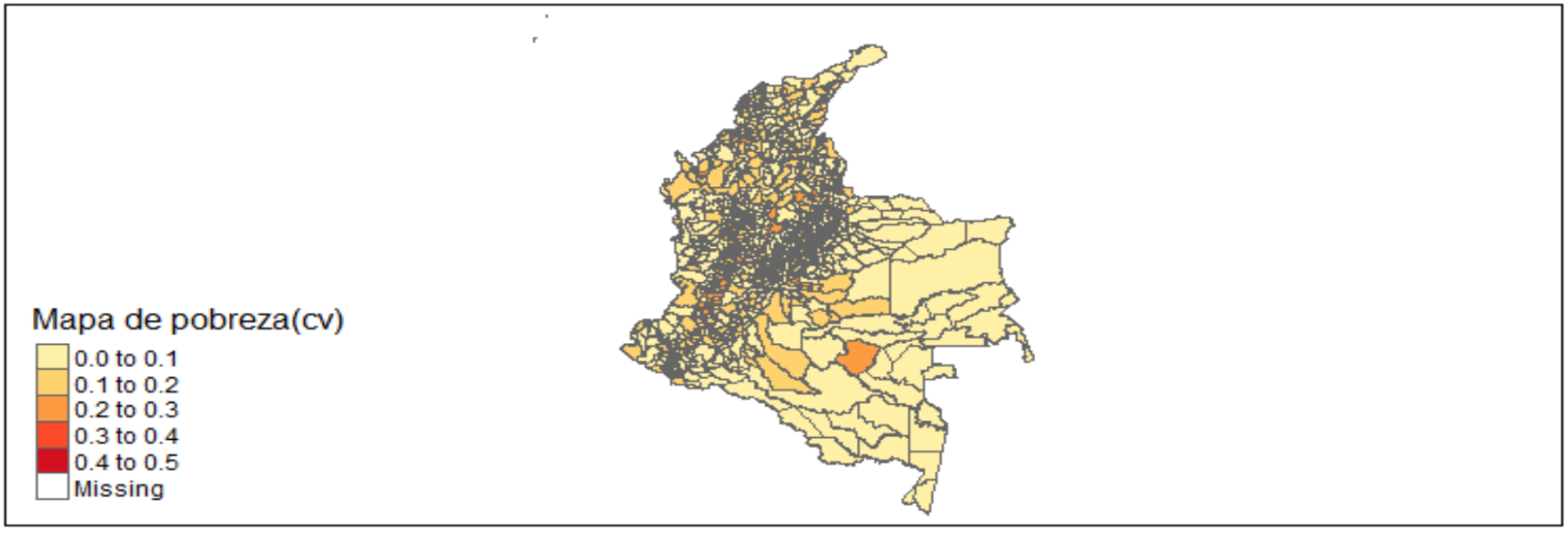


Figure 20: Mapa de los coeficientes de variación

¡Gracias!

Email: andres.gutierrez@cepal.org