PONTIFICIA UNIVERSIDAD CATÓLICA DEL PERÚ

Escuela de Posgrado

Estimación de áreas pequeñas mediante modelos aditivos de ubicación, escala y forma aplicados a una encuesta de hogares en Perú

Tesis para obtener el grado académico de Maestro en Estadística que presenta:

Hans Stehli Torrecilla

Asesor:

Luis Hilmar Valdivieso Serrano

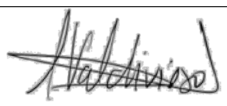
Lima, 2024

Informe de Similitud

Yo Luis Hilmar Valdivieso Serrano docente de la Escuela de Posgrado de la Pontificia Universidad Católica del Perú, asesor de la tesis titulada "Estimación de áreas pequeñas mediante modelos aditivos de ubicación, es-cala y forma aplicados a una encuesta de hogares en Perú", del autor Stehli Torrecilla, Hans dejo constancia de lo siguiente:

- El mencionado documento tiene un índice de puntuación de similitud de 7 %. Así lo consigna el reporte de similitud emitido por el software Turnitin el 08/07/2024.
- He revisado con detalle dicho reporte y confirmo que cada una de las coincidencias detectadas no constituyen plagio alguno.
- Las citas a otros autores y sus respectivas referencias cumplen con las pautas académicas.

Lima, 09 de Julio de 2024

Apellidos y nombres del asesor: Valdivieso Serrano, Luis Hilmar	
DNI: 07958730 ORCID: https://orcid.org/0000-0002-8975-7557	Firma: 

Resumen

El objetivo de la presente tesis es evaluar la robustez de los modelos aditivos de ubicación, escala y forma (GAMLSS) en una estimación en áreas pequeñas. Para ello, se realizan simulaciones estadísticas en donde se aplican estos modelos para diferentes distribuciones de la variable dependiente considerando distintos niveles de variabilidad entre las áreas, analizando la precisión de los resultados en cada caso. Asimismo, se realiza una aplicación utilizando la Encuesta Nacional de Hogares de Perú (ENAHO) del año 2017 para obtener indicadores de infraestructura de hogares y sus intervalos de confianza a nivel distrital para el departamento de Ica, además de contrastar las estimaciones con las cifras poblacionales obtenidas del Censo Nacional del mismo año.

Los resultados revelan que los indicadores obtenidos mediante GAMLSS tienen un menor error cuadrático medio que aquellos estimados de manera directa, considerando el diseño muestral. Asimismo, se encuentra que los GAMLSS generan resultados más exactos respecto a los valores poblacionales, aunque ello depende de la heterogeneidad de las áreas. Este hallazgo es consistente aún bajo el supuesto de una variable dependiente de tipo dicotómica (balanceada o no balanceada) o de tipo numérica (discreta o continua). Asimismo, estas bondades son más evidentes si el tamaño de las muestras de las áreas es reducido. Finalmente, a través de la aplicación, se han obtenido estimaciones puntuales y intervalos de confianza para indicadores de acceso a saneamiento y número de habitaciones de las viviendas, correspondientes a 37 distritos del departamento de Ica.

Palabras clave: estimación en áreas pequeñas, GAMLSS, indicadores, distritos, saneamiento, infraestructura, viviendas, ENAHO

Abstract

The objective of this thesis is to evaluate the robustness of additive models of location, scale and shape (GAMLSS), under the framework of small area estimation. To do this, statistical simulations are carried out for different distributions of the dependent variable, considering different levels of variability between areas and analyzing the precision of the results in each case. Likewise, an application is carried out using the National Household Survey of Peru (ENAHO) from 2017 to obtain household infrastructure indicators and their confidence intervals at the district level for the department of Ica. These estimations were later compared to their population values, retrieved from the National Census of the same year.

The results reveal that the indicators obtained through GAMLSS have a lower mean square error than those estimated directly by considering the sample design. Likewise, it is found that the GAMLSS generate more accurate results with respect to population values, although this depends on the heterogeneity of the areas. This finding is consistent even under the assumption of a dichotomous (balanced or unbalanced) or numerical (discrete or continuous) dependent variable. Likewise, these benefits are more evident if the sample size of the areas is small. Finally, through the application, point estimates and confidence intervals have been obtained for indicators of access to sanitation and number of rooms of the households, corresponding to 37 districts of the department of Ica.

Keywords: Small area estimation, GAMLSS, indicators, districts, sanitation, infrastructure, households, ENAHO

Índice

Índice de tablas

Índice de figuras

Capítulo 1
Introducción

1.1. Consideraciones preliminares

El análisis de información es el primer paso para una toma de decisiones acertada y eficaz. Los datos permiten comprender las dimensiones de alguna situación de interés, así como su magnitud y urgencia. En ese sentido, las entidades tanto públicas como privadas utilizan información de manera continua para priorizar y planificar sus próximas acciones. Respecto a la producción de estadísticas oficiales en Perú, el Instituto Nacional de Estadística e Informática (INEI) es el órgano responsable de la "producción y difusión de información estadística en forma oportuna, y confiable, para el mejor conocimiento de la realidad nacional y la adecuada toma de decisiones"[1]. Para ello, recurre a censos, encuestas e información administrativa.

La mayoría de las estadísticas oficiales de periodicidad anual se estiman a través de encuestas por muestreo como la Encuesta Nacional de Hogares, la Encuesta Permanente de Empleo, la Encuesta Nacional de Programas Presupuestas, entre otras, las cuales proveen datos como las condiciones de los hogares y de su población residente. De acuerdo a sus fichas técnicas, las muestras de estas encuestas permiten obtener estimaciones y realizar inferencia como máximo a un nivel departamental.

En el caso de la Encuesta Nacional de Hogares (ENAHO), su diseño muestral y el tamaño de su muestra (36 996 viviendas particulares) permite obtener estimaciones anuales para un nivel de inferencia nacional, y como máximo para los 24 departamentos y el área metropolitana de Lima y Callao. Para estimaciones de periodicidad trimestral, la representatividad de la muestra solo permite obtener estimaciones confiables a nivel nacional (permitiendo la desagregación entre ámbitos urbano y rural)[2].

Por consiguiente, la muestra tomada no es suficientemente representativa para obtener directamente una estimación confiable o eficiente a nivel provincial o distrital, lo cual representa un obstáculo para la gestión de las autoridades locales quienes solo cuentan con información de los censos nacionales con una periodicidad de cada diez años.

[1] Reglamento de Organización de Funciones del Instituto Nacional de Estadística e Informática.

[2] Ficha técnica de la Encuesta Nacional de Hogares sobre Condiciones de Vida y Pobreza - 2017.

1.2. Objetivos

El objetivo general de la tesis es estudiar las bondades de los modelos aditivos de ubicación, escala y forma (GAMLSS) en un contexto de estimación en áreas pequeñas y realizar una aplicación con datos reales en el Perú con el propósito de aportar en la generación de estadísticas para la toma de decisiones a nivel local. Siguiendo ello, los objetivos específicos son:

- Realizar un estudio sobre el estado del arte de los modelos de estimación en áreas pequeñas y los GAMLSS (incluyendo su aplicación para estimar parámetros para subpoblaciones a partir de muestras reducidas). Algunos trabajos a destacar son los de Rao y Hussain Chounhry (1995), Stasinopoulos y Rigby (2007), Molina y Rao (2010), y Mori y Ferrante (2023).
- Evaluar la robustez de los estimadores GAMLSS en un contexto de estimación en áreas pequeñas mediante simulaciones, considerando diferentes distribuciones para la variable dependiente y niveles de variabilidad entre las áreas.
- Implementar un GAMLSS para estimar y obtener intervalos de confianza para indicadores de infraestructura de los hogares a nivel distrital, utilizando datos de la Encuesta Nacional de Hogares del 2017.
- Evaluar el rendimiento de la predicción de los indicadores mediante la estimación por $bootstrap$ paramétrico del error cuadrático medio utilizada por Mori y Ferrante (2023), así como contrastando las estimaciones con cifras del Censo Nacional del 2017.

1.3. Organización de la tesis

En el capítulo 2, se realizará una revisión de literatura en la que se describe el estado del arte de la estimación en áreas pequeñas y de los GAMLSS. En el capítulo 3, se abordará el marco teórico del modelo propuesto en la presente tesis, indicando las ecuaciones y propiedades más relevantes, así como la explicación del método de estimación. En capítulos posteriores se realizará un estudio de simulación para estudiar las bondades del modelo GAMLSS en un contexto de estimación en áreas pequeñas, así como su aplicación en datos reales usando la Encuesta Nacional de Hogares. La tesis culmina con la descripción de las conclusiones principales, la bibliografía consultada y un apéndice con los códigos del programa estadístico R usados para la aplicación con datos reales.

Capítulo 2
Estimación en áreas pequeñas y GAMLSS

En esta sección se presentan trabajos recientes sobre estimación en áreas pequeñas y los modelos aditivos generalizados de localización, escala y forma (GAMLSS) que sirvan como referencia para plantear el modelo sobre el cual trabajaremos en la presente tesis. Asimismo, se describe la técnica de estimación *bootstrap* que se usará para la implementación del modelo.

2.1. Estimación en áreas pequeñas

Los métodos de estimación en áreas pequeñas tienen la ventaja de permitir obtener la estimación de una característica particular para aquellos casos en los que el tamaño de las muestras en una subpoblación, grupo o región (por ejemplo, un distrito), al cual llamaremos área, no es lo suficientemente elevado o suficiente para asegurar errores estándar aceptables.

Los estimadores de áreas pequeñas pueden ser directos, es decir funciones de la información muestral disponible; sintéticos, utilizando información suplementaria conocida del área; compuestos, es decir un promedio ponderado de los anteriores tipos; o basados en modelos[3]. El modelo más simple para obtener estimadores en áreas pequeñas es descrito por Rao y Hussain Chounhry (1995) de la siguiente manera:

$$y_{ij} = x_{ij}\beta + \alpha_j + \epsilon_{ij}, \quad \begin{matrix} i = 1, 2, ..., n_j \\ j = 1, 2, ..., J \end{matrix} \tag{1}$$

donde y_{ij} representa la variable dependiente de interés de la unidad i en el área j, x_{ij} es un vector de variables auxiliares asociadas a y_{ij}, α_j es un efecto aleatorio diferente para cada área con varianza σ^2_α y ϵ_{ij} es un error aleatorio con media cero y varianza σ^2_ϵ, el cual es independiente de α_j. Nótese además que esta especificación es un caso particular del modelo de interceptos aleatorios.

Se han planteado diversas mejoras a este modelo simple en un contexto de áreas pequeñas. Por ejemplo, el modelo de Fay y Herriot (1979) considera una ecuación similar a (1) pero con datos agregados a nivel de área, cuya específicación se muestra a continuación:

$$y_j = x_j\beta + \alpha_j + e_j, \quad j = 1, 2, ..., J \tag{2}$$

[3]Para mayor información sobre la tipología de estimadores en áreas pequeñas, revisar Rao y Hussain Chounhry (1995).

donde e_j representa un error de muestreo con media cero y varianza σ_j^2 asociado al estimador directo y_j . En este caso, el uso de datos a nivel de área podría ser ventajoso ya que son menos afectado por observaciones extremas presentes en los microdatos (Guadarrama, Molina y Rao, 2014).

Asimismo, si bien el modelo básico asume efectos aleatorios independientes entre áreas, se han desarrollado modelos Fay-Herriot incorporando autocorrelación espacial, permitiendo que la ubicación del área y sus zonas vecinas sean determinantes en la estimación. Recientemente, Haro (2022) aplicó el modelo Fay-Herriot espacial para obtener estimaciones de indicadores de salud para distritos con mayor nivel de pobreza en Perú, usando datos de la Encuesta Demográfica y de Salud Familiar y el Censo Nacional 2017.

Otra mejora del modelo (1) es la regresión lineal de error anidado planteada por Battese, Harter y Fuller (1988). A diferencia del modelo de Fay-Herriot, este permite utilizar covariables tanto a nivel de área como a nivel de unidades secundarias a través de microdatos. Bajo esta especificación, la predicción de los indicadores a nivel de área se obtiene mediante la siguiente expresión:

$$\hat{H}_j = \bar{x}_j^p \tilde{\beta} + \xi_j(\bar{y}_j - \bar{x}_j \tilde{\beta}) \tag{3}$$

donde $\tilde{\beta}$ corresponde al mejor estimador linear insesgado de β del modelo (1) estimado por mínimos cuadrados generalizados o máxima verosimilitud. Por su lado, $\bar{y}_j = \frac{1}{n_j}\sum_{i=1}^{n_j} y_{ij}$ y $\bar{x}_j = \frac{1}{n_j}\sum_{i=1}^{n_j} x_{ij}$ corresponden a las medias muestrales de la variable dependiente y las covariables, respectivamente. Además, $\bar{x}_j^p$ corresponde a las medias poblacionales de las variables auxiliares mientras que $\xi_j = \frac{\tilde{\sigma}_\alpha^2}{\tilde{\sigma}_\alpha^2 + \tilde{\sigma}_\epsilon^2/n_j}$.

El estimador (3) es denominado como el mejor predictor lineal insesgado empírico (EBLUP por sus siglas en inglés) de indicadores a nivel de área. Una de sus principales ventajas es que permite compensar las estimaciones cuando los tamaños de muestra son pequeños. Si n_j es bajo, el cálculo de $\hat{H}_j$ se basará principalmente en $\bar{x}_j^p \tilde{\beta}$; a medida que n_j aumente, la estimación irá incorporando el componente $\xi_j(\bar{y}_j - \bar{x}_j \tilde{\beta})$.

Por otro lado, el modelo (1) ha sido extendido por Elbers, Lanjouw y Lanjouw (2003) al permitir la posibilidad de que los efectos aleatorios α_j correspondan a clusters que no necesariamente equivalgan a áreas. Asimismo, el método incorpora una simulación de los datos poblacionales de la variable dependiente a partir de muestreos con reemplazo de las predicciones $\tilde{y}_{ij} = \bar{x}_{ij}\tilde{\beta} + \tilde{\alpha}_j + \tilde{\epsilon}_{ij}$, donde $\tilde{\alpha}_j \sim N(0, \tilde{\sigma}_\alpha^2)$ y $\tilde{\epsilon}_j \sim N(0, \tilde{\sigma}_\epsilon^2)$, donde $\tilde{\beta}$, $\tilde{\sigma}_\alpha^2$ y $\tilde{\sigma}_\epsilon^2$ son estimados de la misma manera que bajo el modelo de regresión lineal de error anidado. En la actualidad, el Banco Mundial utiliza este modelo para calcular estimadores de pobreza y desigualdad de ingresos para áreas pequeñas en algunos países (Molina y Rao, 2010).

Sin embargo, un inconveniente del modelo de regresión lineal de error anidado es que asume normalidad e independencia de los efectos aleatorios α_j y los errores ϵ_{ij}. De hecho, según Rojas-Perilla, Pannier, Schmid y Tzavidis (2020) si el supuesto de normalidad de α_j y ϵ_{ij} no se cumple, entonces las predicciones del modelo incurrirán en sesgos. Para remediar ello, estos autores han propuesto el uso de diversas transformaciones de la variable dependiente continua como Log-Shift, Box-Cox, y Dual Power para facilitar que los errores tengan una distribución aproximadamente normal. De esta manera, los autores demuestran que transformar la variable dependiente aumenta la capacidad predictiva y permite abordar en mejor medida los supuestos del modelo; empero se evidencia que los residuos a nivel de unidades secundarias no llegan a distribuirse de manera normal nisiquiera con las transformaciones utilizadas.

La linealidad del modelo de Battese, Harter y Fuller (1988) es otro de sus inconvenientes, ya que no permite incorporar funciones de enlace en el modelamiento de las variables dependientes. Ello es de particular importancia si la variable de interés es de tipo binaria. Finalmente, la aplicación del modelo puede resultar impráctica ya que requiere las medias poblacionales de las covariables en cada área, cuya información no necesariamente se tenga disponible periódicamente.

2.2. Modelos aditivos de ubicación, escala y forma (GAMLSS)

Los modelos aditivos generalizados de ubicación, escala y forma (GAMLSS por sus siglas en inglés) son un enfoque relativamente moderno para la estimación de regresiones de manera semi-paramétrica. Planteados por Stasinopoulos y Rigby (2007), consisten en la estimación de los parámetros de la distribución de una variable dependiente $f(y \mid \boldsymbol{\theta})$, no necesariamente de naturaleza continua ni normal, como función de covariables y una suma efectos aleatorios:

$$g_k(\theta_k) = \mathbf{X}^k \boldsymbol{\beta}_k + \sum_{m=1}^{M_k} \mathbf{Z}_m^k \boldsymbol{\gamma}_m^k, \quad k = 1, 2, 3, 4 \tag{4}$$

donde $g_k(\cdot)$ es una función de enlace[4], θ_k es una componente del vector de parámetros $\boldsymbol{\theta}$ asociado a la función de masa de la variable dependiente y, $\mathbf{X}^k$ es una matriz de covariables, y $\sum_{m=1}^{M_k} \mathbf{Z}_m^k \boldsymbol{\gamma}_m^k$ representan M_k términos aditivos correspondientes a los efectos aleatorios que son asignados a las observaciones mediante las matrices de diseño $\mathbf{Z}_m^k$ (conocidas). En particular, $\boldsymbol{\theta} = (\mu, \sigma, \nu, \tau)$ usualmente hacen referencia a los parámetros de

[4]La elección de la función de enlace $g_k(\cdot)$ dependerá de la distribución elegida para modelar la variable dependient $f(y \mid \boldsymbol{\theta})$. Por ejemplo, para variables numéricas continuas normales, la función de enlace puede ser la función identidad, mientras que para variables numéricas con distribución $Gamma$, la función de enlace puede ser una logarítmica.

ubicación (media), escala (desviación estándar) y forma (asimetria y curtosis) de $f(y \mid \boldsymbol{\theta})$, sin embargo, los GAMLSS pueden adecuarse a la forma de cualquier distribución y por ende admiten la modelación de más de cuatro parámetros.

Nóteses que la cualidad "semi-paramétrica"de los GAMLSS proviene de dos características: el modelo asume una forma parámetrica para la distribución de la variable dependiente $f(y \mid \boldsymbol{\theta})$ y, a la vez, incorpora términos aditivos que implican el uso de funciones no paramétricas de suavización, de manera similar a los modelos aditivos generales (GAM).

La principal ventaja de estos modelos es la flexibilidad que ofrecen para modelar la distribución de la variable dependiente, lo cual es de particular utilidad cuando los datos presentan mucha asimetría o curtosis, o cuando no se ajustan a una distribución de la familia exponencial. Desde su implementación en el programa estadístico R, los GAMLSS han sido aplicados en diversas áreas como las finanzas, medicina, meteorología, consumo de bienes y servicios, entre otros; asimismo es aplicado por instituciones conocidas como el Fondo Monetario International y la Organización Mundial de la Salud (Stasinopoulos, Rigby y Bastiani, 2018).

Un ejemplo que ilustra las bondades de la flexibilidad de los GAMLSS es el trabajo de Voudouris, et al. (2012), en el que se aplica este tipo de modelos para aproximar un conjunto de datos de ingresos de películas, cuya principal característica es su elevada asimetria y curtosis. Utilizando una distribución exponencial Box-Cox de cuatro parámetros, los autores encuentran que el GAMLSS realiza mejores predicciones de los datos, respecto a otros modelos, como los basados en la distribución Pareto-Levy-Mandelbro, utilizados en el pasado para modelar los ganacias de taquilla.

Otro ejemplo de aplicación es el trabajo de López y Francés (2013) en el cual se aprovechan los GAMLSS para modelar series de tiempo no estacionarias de frecuencias de inundaciones usando variables explicativas meteorológicas. En este caso, se encuentra que la flexibilidad de los GAMLSS permite describir de mejor manera la no estacionaridad de las series de tiempo a diferencia de modelos tendenciales simples; asimismo aportan en la predicción de inundaciones futuras debido al uso de covariables en la estimación.

Los GAMLSS han sido usados recientemente en el ámbito de la estimación en áreas pequeñas por Mori y Ferrante (2023), quienes demuestran mediante simulaciones y estimación por $bootstrap$ que las predicciones a nivel de área (indicadores) con este método tienen una varianza similar o menor a la obtenida con el modelo de regresion lineal de errores anidados de Battese, Harter y Fuller (1988). Asimismo, aplican el modelo a datos reales para estimar el gasto per cápita de la población de Italia a nivel regional a pesar de tener tamaños de muestra reducidos para algunas zonas.

La estimación de los parámetros $\boldsymbol{\beta}_k$ y de los efectos aleatorios $\boldsymbol{\gamma}_m^k$ para $k = 1, 2, 3, 4$ y $m = 1, \ldots, M_k$ bajo un GAMLSS se realiza a través de la maximización de la siguiente función de log-verosimilitud penalizada:

$$\ell_p = \sum_{i=1}^{n} log[f(y_i \mid \boldsymbol{\theta}_i)] - \frac{1}{2} \sum_{k=1}^{4} \sum_{m=1}^{M_k} \lambda_m^k \gamma_m^{k\top} G_m^k \gamma_m^k \tag{5}$$

donde G_m^k representa la inversa de la matriz varianza-covarianza de γ_m^k cuya especificación puede depender de un vector de hiperparámetros λ_m^k. En el caso que (4) no tuviera términos aditivos, la función de log-verosimilud penalizada se reduciría a una función de log-verosimilud estándar (Stasinopoulos y Rigby, 2007).

La función de verosimilitud penalizada (5) se puede maximizar aplicado diferentes algoritmos. El primero es el algoritmo de Cole y Green que utiliza la primera y segunda derivada, así como las derivadas cruzadas, de la función de verosimilitud con respecto a los parámetros $\boldsymbol{\theta} = (\mu, \sigma, \nu, \tau)$. El segundo es un algoritmo de optimización para el ajuste modelos aditivos de media y dispersión[5].

El comando $gamlss$ en el programa R permite realizar esta estimación considerando uno o ambos algoritmos. Asimismo, el comando permite incorporar pesos dentro de la log-verosimilitud (verosimilitud ponderada) tal que la contribución de cada observación en la ecuación (5) sea diferente. Esto es ventajoso cuando las observaciones han sido muestreadas bajo un esquema multietápico de conglomerados o de estratificación, donde las probabilidades de selección son distintas entre las unidades.

2.3. Estimaciones mediante bootstrap

El método *bootstrap* fue propuesto por Efron (1979) como una técnica para estimar la distribución muestral de una estadística a partir de datos observados en una muestra. El autor encuentra que dicho método resulta más confiable y aplicable que otras técnicas, como *jackknife*, para estimar el sesgo y la varianza de algún estadístico de interés.

Bootstrap no paramétrico

De acuerdo a Efron y Tibshirani (1993), el proceso general *bootstrap* para aproximar de manera no parámetrica la distribución muestral de una estadística $S(\boldsymbol{X})$ a partir de una muestra de tamaño n con observaciones $\boldsymbol{x} = (x_1, x_2, ..., x_n)$ sería el siguiente:

[5]Para mayor información sobre los algoritmos y la maximización de la verosimilitud penalizada, revisar Stasinopoulos y Rigby (2007) y Rigby y Stasinopoulos (1996).

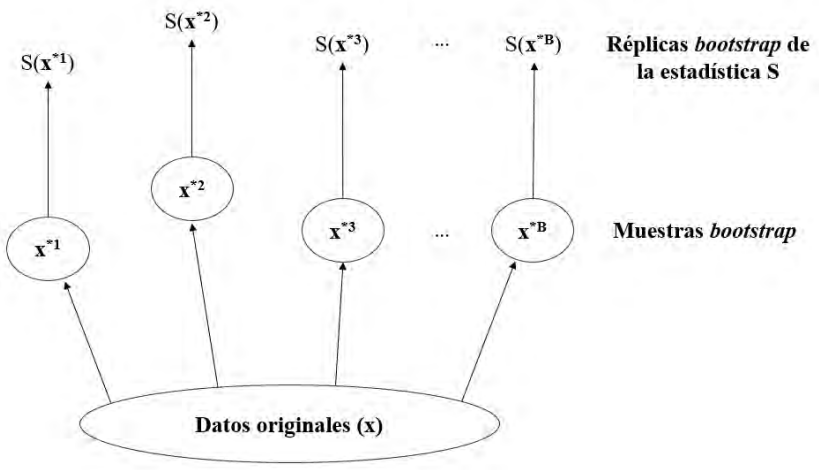


Figura 1: Representación gráfica del método *bootstrap* no paramétrico.
Imagen adaptada de Efron y Tibshirani (1993).

1. Realizar un muestreo con reemplazamiento para los n datos disponibles. Esta nueva muestra $\boldsymbol{x}^{*1}$ se denomina "muestra *bootstrap*".
2. Calcular el estadístico de interés $S(\boldsymbol{x}^{*1})$ con los datos de la muestra *bootstrap*.
3. Repetir los pasos 1 y 2 B veces (se recomienda entre 50 y 200), de tal manera que se tengan al final del proceso B réplicas *bootstrap* de la estadística de interés: $S(\boldsymbol{x}^{*1}), S(\boldsymbol{x}^{*2}), \ldots, S(\boldsymbol{x}^{*B})$.

Con estas réplicas, se puede obtener una aproximación de la distribución muestral de $S(\boldsymbol{x})$ mediante su histograma, o calcular resúmenes como su media $\overline{S(\boldsymbol{x})} = \sum_{b=1}^{B} S(\boldsymbol{x}^{*b})/B$ y su error estándar $se(S(\boldsymbol{x})) = \sqrt{\sum_{b=1}^{B}(S(\boldsymbol{x}^{*b}) - \overline{S(\boldsymbol{x})})^2/(B-1)}$. Nótese que durante el proceso descrito líneas arriba, no se ha realizado ningún supuesto sobre el proceso generador de datos de $\boldsymbol{X}$, ya que se parte de su distribución empírica denotada por sus observaciones $(x_1, x_2, ..., x_n)$. Esta es la principal ventaja del *bootstrap* no paramétrico.

Bootstrap paramétrico

Una manera alternativa de abordar la distribución muestral del estadístico $S(\boldsymbol{X})$ es a través de la simulación de datos $\boldsymbol{X}$ asumiendo que su proceso generador de datos sigue una distribución conocida $f(\boldsymbol{X} \mid \boldsymbol{\theta})$. El proceso para realizar el *bootstrap* paramétrico es similar al no parámetrico, siendo diferente la manera en la que se obtienen las muestras *bootstrap* $\boldsymbol{x}^*$. A continuación, se describe brevemente:

1. Ajustar los datos de la muestra $\boldsymbol{x} = (x_1, x_2, ..., x_n)$ a la distribución $f(\boldsymbol{X} \mid \boldsymbol{\theta})$ y obtener el estimador $\hat{\boldsymbol{\theta}}$.

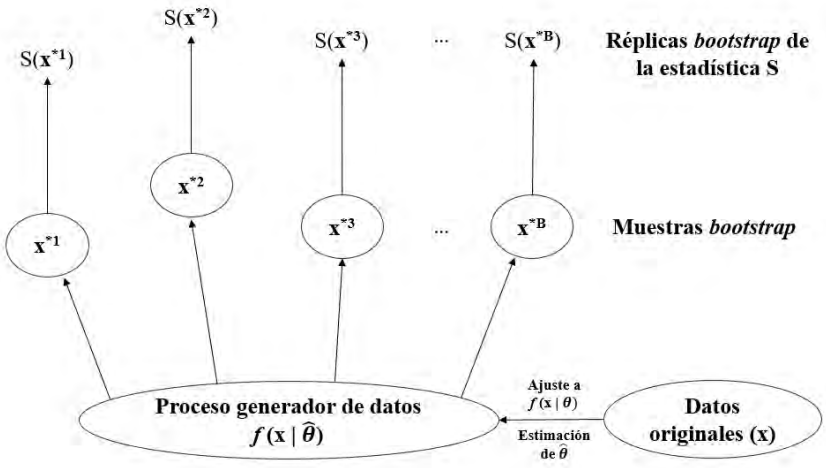


Figura 2: Representación gráfica del método *bootstrap* paramétrico.
Imagen adaptada de Efron y Tibshirani (1993).

2. Obtener una muestra *bootstrap* $\boldsymbol{x}^{*1}$ de tamaño n a partir de la simulación de datos sintéticos de $f(\boldsymbol{X} \mid \hat{\boldsymbol{\theta}})$.

3. Calcular el estadístico de interés $S(\boldsymbol{x}^{*1})$ con los datos de la muestra *bootstrap*.

4. Repetir los pasos 2 y 3 B veces, de tal manera que se tengan al final del proceso B réplicas *bootstrap* de la estadística de interés.

De la misma manera que con el *bootstrap* no paramétrico, las réplicas $S(\boldsymbol{x}^{*1}), \ldots, S(\boldsymbol{x}^{*B})$ permiten calcular sus resúmenes respectivos. Además, es posible obtener los percentiles $P_{0.025}$ y $P_{0.975}$ que corresponden a los límites superiores e inferiores del intervalo de confianza al 95 % de $S(\boldsymbol{X})$. Alternativamente, se pueden plantear supuestos sobre la forma de la distribución muestral de $S(\boldsymbol{X})$ y construir los intervalos de confianza con su error estándar.

Según Efron (1979), las estimaciones realizadas mediante *bootstrap* paramétrico y no paramétrico tenderán a resultados simulares a medida que el tamaño de muestra aumente y si el supuesto $\boldsymbol{X} \sim f(\boldsymbol{X} \mid \boldsymbol{\theta})$ es válido. En ese sentido, la aplicación de *bootstrap* paramétrico en el marco de los GAMLSS es más conveniente, ya que dichos modelos se enfocan en aproximar la distribución de la variable dependiente.

Capítulo 3
Estimación en áreas pequeñas mediante un GAMLSS

En el siguiente capítulo, se presenta la especificación de los GAMLSS para la estimación en áreas pequeñas y la descripción de sus bondades en este contexto. Asimismo, se describen los procesos para la estimación de las estadísticas de interés y de su error cuadrático medio mediante simulaciones por el método *bootstrap*.

3.1. Especificación del modelo

Siguiendo el trabajo de Mori y Ferrante (2023), se desarrolla a continuación el marco teórico de aplicación de los GAMLSS en un contexto de estimación en áreas pequeñas. Considerando una variable dependiente Y_{ij} para la unidad i en el área j, proveniente de una muestra de tamaño n de una población de tamaño N que se puede dividir en J subpoblaciones o áreas de tamaños N_j, se plantea el siguiente GAMLSS para la estimación en áreas pequeñas:

$$g_k(\theta_k) = \mathbf{X}^k\boldsymbol{\beta}_k + \mathbf{Z}^k\boldsymbol{\gamma}^k, \quad k = 1, 2, 3, 4$$

Nótese que se asume un solo término aditivo ($M_k = 1$) que representa el efecto aleatorio de las áreas $\boldsymbol{\gamma}^k_{J\times 1}$ que son asignadas a cada unidad mediante la matriz de diseño $\mathbf{Z}^k_{n\times J}$ según su ubicación geográfica. La versión desagregada de este modelo según unidad i, área j y parámetro k sería:

$$\begin{aligned} g_\mu(\mu_{ij}) &= \mathbf{X}^\mu_{ij}\boldsymbol{\beta}_\mu + \gamma^\mu_j \\ g_\sigma(\sigma_{ij}) &= \mathbf{X}^\sigma_{ij}\boldsymbol{\beta}_\sigma + \gamma^\sigma_j \\ g_\nu(\nu_{ij}) &= \mathbf{X}^\nu_{ij}\boldsymbol{\beta}_\nu + \gamma^\nu_j \\ g_\tau(\tau_{ij}) &= \mathbf{X}^\tau_{ij}\boldsymbol{\beta}_\tau + \gamma^\tau_j \end{aligned} \tag{6}$$

Bajo esta especificación, se asume normalidad e independencia de los efectos aleatorios, es decir, $\gamma^k_j \overset{iid}{\sim} N(0, \boldsymbol{\Psi}_k)$ para $k = \mu, \sigma, \nu, \tau$; donde $\boldsymbol{\Psi}_{k_{(J\times J)}}$ representa una matriz varianza covarianza diagonal con elementos σ^2_k, que puede simplificarse a $\sigma^2_k\mathbf{I}_{(J\times J)}$. El hecho que los efectos aleatorios de las áreas sean independientes, implica suponer que la estimación de los parámetros de la distribución en una ubicación no es influenciada por sus zonas vecinas. En ese sentido, el modelo asume ausencia de autocorrelación espacial al momento de estimar el modelo (6).

Asimismo, la función de masa condicional de la variable dependiente viene dada por $Y_{ij}|\gamma^\mu_j, \gamma^\sigma_j, \gamma^\nu_j, \gamma^\tau_j \sim F(\cdot \mid \mu_{ij}, \sigma_{ij}, \nu_{ij}, \tau_{ij})$ para toda observación i dentro de las áreas $j = 1, ..., J$, siendo F una función de distribución adecuada para la naturaleza de la variable de

interés Y_{ij}. Por ejemplo, en caso Y_{ij} sea una variable dicotómica con valores posibles 0 y 1, entonces F será la función de distribución de una variable aleatoria $Bernoulli(\pi_{ij})$ con un único parámetro $\pi_{ij} = \mu_{ij}$. En caso Y_{ij} sea una variable continua con valores mayores a cero, entonces F se podría plantear con una distribución $Gamma(\alpha_{ij}, \beta_{ij})$ con parámetros $\alpha_{ij} = 1/\sigma_{ij}^2$ y $\beta_{ij} = 1/(\mu_{ij}\sigma_{ij}^2)$.[6]

Según la especificación anterior, los GAMLSS se diferencian de otros modelos de estimación en áreas pequeñas principalmente por la forma en la que se aborda la variable dependiente. En lugar de modelar Y_{ij} con una ecuación basada en covariables y errores aleatorios, se modelan los parámetros de la función de masa de Y_{ij}, la cual puede no ser necesariamente normal e incluso podría no pertenecer a la familia exponencial. Esta flexibilidad permite relajar el supuesto típico de normalidad de errores en una estimación en áreas pequeñas, como es el caso del modelo (1) o el de Fay-Herriot y Battese, Harter y Fuller.

Si se asume que la función de distribución $F(\cdot \mid \mu_{ij}, \sigma_{ij}, \nu_{ij}, \tau_{ij})$ es correcta para caracterizar la variable dependiente en toda la población, entonces será posible obtener una estimación de los parámetros de la ditribución F usando las observaciones de la muestra de tamaño n. Para obtener las estimaciones de los parámetros $\delta_{ij} = (\mu_{ij}, \sigma_{ij}, \nu_{ij}, \tau_{ij})$ así como de $\boldsymbol{\beta} = (\boldsymbol{\beta}_\mu, \boldsymbol{\beta}_\sigma \boldsymbol{\beta}_\nu, \boldsymbol{\beta}_\tau)$ y $\boldsymbol{\gamma}_j = (\gamma_j^\mu, \gamma_j^\sigma, \gamma_j^\nu, \gamma_j^\tau)$, se aplica la metodología descrita en la sección 2.2.

Por otro lado, si bien la especificación del modelo (6) no tiene una forma explícita de residuos, es posible obtener los denominados residuos aleatorios cuantílicos (*randomized quantile residuals*) mediante:

$$r_{ij} = \Phi^{-1}(\mathcal{F}(y_{ij} \mid \hat{\mu}_{ij}, \hat{\sigma}_{ij}, \hat{\nu}_{ij}, \hat{\tau}_{ij}))$$

donde Φ^{-1} representa la inversa de la función de densidad acumulada de una distribución normal estándar y $\mathcal{F}$ representa la distribución acumulada de $F(\cdot \mid \hat{\mu}_{ij}, \hat{\sigma}_{ij}, \hat{\nu}_{ij}, \hat{\tau}_{ij})$. De acuerdo a Dunn y Smyth (1996), estos residuos son exactamente normales si los parámetros de la distribución ajustada son estimados consistentemente. Asimismo, son particularmente útiles para realizar diagnósticos de calidad de ajuste de modelos si la variable dependiente de interés no es continua. Para estos casos, se introduce un componente de aleatoriedad uniforme para evitar la superposición de las observaciones.

[6] Para mayor información sobre este tipo de reparametrización, ver sección 4.2.

3.2. Estimación de características de áreas pequeñas

El principal interés en el uso de los GAMLSS en un contexto de estimación en áreas pequeñas es obtener estimaciones de alguna característica vinculada a la variable dependiente $\boldsymbol{Y}$ para las áreas. En términos generales, se desea una estimación de $H_j = \zeta(\boldsymbol{Y}_j)$ que puede ser una media, proporción o cualquier función de los N_j datos de la variable dependiente en el área j, en base a los datos muestrales. Si la distribución de la variable dependiente fuese conocida, entonces será de interés estimar:

$$\tilde{H}_j = E(\zeta(\boldsymbol{Y}_j)) = \int \zeta(y) f_j(y) dy, \quad \text{para} \quad j = 1, 2, ..., J \tag{7}$$

que es un parámetro de la superpoblación[7] del área j. Por su lado, $\tilde{H}_j$ es una estimación asociada a un proceso generador de datos hipotético, mientras que H_j es una estadística obtenida mediante datos finitos. En ese sentido, si el proceso generador de datos de la superpoblación es modelado idóneamente entonces $\tilde{H}_j \approx H_j$.

No obstante, debe tomarse en cuenta que la función de masa $f_j(y)$, asociada a la superpoblación de $\boldsymbol{Y}$, no es conocida y que solo se cuenta con una muestra que contiene observaciones limitadas de las areas $j = 1, \ldots, J$. En ese sentido, Graf, Marin y Molina (2019) y Mori y Ferrante (2023) proponen una simulación de Monte Carlo para aproximar $\tilde{H}_j$ usando los datos de la muestra de tamaño n y la función de distribución condicional de la variable dependiente $\hat{F}(\cdot \mid \hat{\mu}_{ij}, \hat{\sigma}_{ij}, \hat{\nu}_{ij}, \hat{\tau}_{ij})$, estimada mediante GAMLSS y asumiendo que permite abordar correctamente el proceso estocástico hipotético de la superpoblación de $\boldsymbol{Y}$. La simulación de Monte Carlo se describe a continuación:

1. Estimar (6) usando los datos de la muestra y obtener el vector de estimadores $\hat{\delta}_{ij} = (\hat{\mu}_{ij}, \hat{\sigma}_{ij}, \hat{\nu}_{ij}, \hat{\tau}_{ij})$.
2. Generar datos sintéticos y_j^ℓ de la variable dependiente $\boldsymbol{Y}_j^\ell$ a partir de $\hat{F}(\cdot \mid \hat{\mu}_{ij}, \hat{\sigma}_{ij}, \hat{\nu}_{ij}, \hat{\tau}_{ij})$. Se generará la cantidad de datos sintéticos necesarios para completar los tamaños de las subpoblaciones finitas, es decir, $N_j - n_j \; \forall \; j = 1 \ldots J$.
3. Obtener una estimacion $H_j^\ell = \zeta(y_j^\ell)$ usando los datos de la variable dependiente de la muestra junto con los datos sintéticos generados en el paso anterior.
4. Repetir los pasos 2 y 3 para $\ell = 1, ..., L$, obteniendo finalmente L valores de H_j^ℓ.

[7] La superpoblación es un concepto teórico que alude a la premisa que una población finita observada es en realidad una muestra de una población hipotética infinita, denominada superpoblación (Graubard y Korn, 2002).

La estimación por Monte Carlo de $\tilde{H}_j$ será:

$$\hat{\tilde{H}}_j \approx \frac{1}{L}\sum_{l=1}^{L} H_j^{\ell} \tag{8}$$

Con ello, es posible obtener estimaciones directas de la característica de interés (media, proporción, etc) de la variable dependiente para las áreas $j = 1, \ldots, J$. Adicionalmente, se requiere de una estimación del error cuadrático medio para evaluar la calidad[8] de estos estimadores. Teóricamente, el error cuadrático medio de $\hat{\tilde{H}}_j$ es:

$$MSE(\hat{\tilde{H}}_j) = E((\hat{\tilde{H}}_j - \tilde{H}_j)^2), \quad \text{para} \quad j = 1, 2, ..., J$$

donde el valor esperado se toma con respecto a la distribución del estimador $\hat{\tilde{H}}_j$. De manera similar al cálculo de $\hat{\tilde{H}}_j$, Mori y Ferrante (2023) y Gonzalez-Manteiga, Lombardía, Morales y Santamaría (2008) plantean el siguiente método alternativo por simulación para aproximar el error cuadrático medio de $\hat{\tilde{H}}_j$ a través de una estimación por *bootstrap* paramétrico:

1. Estimar (6) usando los datos de la muestra y obtener los vectores de estimadores $\hat{\delta}_{ij} = (\hat{\mu}_{ij}, \hat{\sigma}_{ij}, \hat{\nu}_{ij}, \hat{\tau}_{ij})$, así como $\hat{\boldsymbol{\beta}}$ y la estimación de las matrices de varianza-covarianza $\hat{\boldsymbol{\Psi}}_k$ para $k = \mu, \sigma, \nu, \tau$.

2. Para cada parámetro μ, σ, ν, τ generar un vector $\mathbf{t}^*_{k_{(J\times 1)}}$ con J realizaciones de una distribución normal estándar y evaluar $\gamma^{k*} = \hat{\sigma}_k \mathbf{I} \times \mathbf{t}^*_k$. Ello es equivante a simular los efectos aleatorios $\gamma_j^k \overset{iid}{\sim} N(0, \hat{\boldsymbol{\Psi}}_k)$ bajo el supuesto de independencia de los errores aleatorios de las áreas.

3. Obtener nuevas estimaciones $bootstrap$ de $\hat{\mu}^*_{ij}, \hat{\sigma}^*_{ij}, \hat{\nu}^*_{ij}, \hat{\tau}^*_{ij}$ mediante:

$$\begin{aligned}
\hat{\mu}^*_{ij} &= g_\mu^{-1}(\mathbf{X}^\mu_{ij}\hat{\boldsymbol{\beta}}_\mu + \mathbf{Z}^\mu_j \boldsymbol{\gamma}^{\mu*}_j) \\
\hat{\sigma}^*_{ij} &= g_\sigma^{-1}(\mathbf{X}^\sigma_{ij}\hat{\boldsymbol{\beta}}_\sigma + \mathbf{Z}^\sigma_j \boldsymbol{\gamma}^{\sigma*}_j) \\
\hat{\nu}^*_{ij} &= g_\nu^{-1}(\mathbf{X}^\nu_{ij}\hat{\boldsymbol{\beta}}_\nu + \mathbf{Z}^\nu_j \boldsymbol{\gamma}^{\nu*}_j) \\
\hat{\tau}^*_{ij} &= g_\tau^{-1}(\mathbf{X}^\tau_{ij}\hat{\boldsymbol{\beta}}_\tau + \mathbf{Z}^\tau_j \boldsymbol{\gamma}^{\tau*}_j)
\end{aligned}$$

4. Generar observaciones sintéticas de la variable dependiente para cada área a través del modelo $\hat{F}(\cdot \mid \hat{\mu}^*_{ij}, \hat{\sigma}^*_{ij}, \hat{\nu}^*_{ij}, \hat{\tau}^*_{ij})$. Este vector sería $\mathbf{y}^*_j = (y^*_{1j}, ..., y^*_{N_j j})$.

5. Obtener $H_j^{*(b)} = \zeta(\mathbf{y}^*_j)$ que sería una estimación de $\tilde{H}_j$.

[8] El error cuadrático medio es una medida de la calidad de un estimador ya que considera tanto su varianza, es decir que tanto se desvían los cálculos bajo diferentes muestras respecto a su promedio, y su sesgo, es decir que tanto se desvia el promedio estimado de su valor poblacional. Por ende, en el caso de estimadores insesgados, su error cuadrático medio equivale a su varianza.

6. Estimar nuevamente el modelo (6) considerando las covariables $\mathbf{X}$ y los datos generados de la variable dependiente $\mathbf{Y}^*$ para aquellos casos contenidos en la muestra, obteniendo una estimación *bootstrap* $\hat{\delta}^*$.

7. Obtener una estimación *bootstrap* de $\hat{\hat{H}}_j^{*(b)}$ considerando el proceso de simulación por Monte Carlo descrito anteriormente para (8).

8. Repetir los pasos (2)-(7) para $b = 1, ..., B$, obteniendo finalmente B valores de $H_j^{*(b)}$ y $\hat{\hat{H}}_j^{*(b)}$.

La estimación por *bootstrap* del error cuadrático medio de $\hat{H}_j$ será:

$$\widehat{MSE}_B(\hat{\hat{H}}_j) = \frac{1}{B}\sum_{b=1}^{B}(\hat{\hat{H}}_j^{*(b)} - H_j^{*(b)})^2 \tag{9}$$

Finalmente, con este error cuadrático medio, es posible obtener una estimación del intervalo de confianza al 100(1-α) % para $\hat{\hat{H}}_j$, asumiendo una forma conocida para su distribución muestral dependiendo de si $\hat{\hat{H}}_j$ es un promedio, proporción, varianza, etc. Por ejemplo, en caso H_j corresponda a una proporción, se podrá asumir una distribución muestral aproximadamente normal, por lo que se utilizaría el cuantil $z_{1-\frac{\alpha}{2}}$ de una distribución normal estándar para construir el intervalo de confianza respectivo.

Observe que los métodos de Monte Carlo y *bootstrap* descritos anteriormente requieren la generación de observaciones sintéticas para toda la población y la estimación del GAMLSS múltiples veces, por lo que el proceso puede ser computacionalmente intensivo si el tamaño de N es elevado (Graf, Marin y Molina, 2019).

Capítulo 4

Evaluación de la robustez de estimadores GAMLSS en un contexto de estimación de áreas pequeñas

En el presente capítulo, se pretende demostrar las propiedades de los estimadores GAMLSS cuando estos se aplican para estimar características en áreas pequeñas, implementando los modelos descritos anteriormente. En ese sentido, se realizarán simulaciones bajo distintos escenarios, comparando la calidad de los estimadores GAMLSS con el de estimadores directos basados solo en el diseño muestral de los datos, así como con los estimadores del modelo de regresión lineal de error anidado de Battese, Harter y Fuller (1988). En particular, se desea evaluar si los estimadores GAMLSS son relativamente más robustos, considerando diferentes tipos de variables dependientes y diferentes niveles de variabilidad entre las áreas.

Para la comparación de los estimadores, se evaluarán tres estadísticas de manera agregada:

- El error cuadrático medio (MSE) promedio, el cual se calculará a partir de las estimaciones a nivel de área por $bootstrap$ en el caso de los GAMLSS de acuerdo a la ecuación (9).

$$\overline{MSE} = \frac{\sum_{j=1}^{J} \widehat{MSE}_B(\hat{\bar{H}}_j)}{J}$$

  En el caso de los estimadores directos basados en el diseño muestral, se asumirá que son insesgados por lo que se considerará su varianza. Respecto a los estimadores de Battese, Harter y Fuller (1988), se estimará su error cuadrático medio a través de un *bootstrap* paramétrico.

- La diferencia relativa promedio respecto al valor poblacional, conocido bajo las simulaciones planteadas.

$$\overline{Diff} = \frac{1}{J} \sum_{j=1}^{J} \frac{|\hat{\bar{H}}_j - H_j|}{H_j}$$

- La proporción de áreas cuyo valor poblacional H_j no se encuentre dentro del intervalo de confianza al 95 % del estimador $\hat{\bar{H}}_j$.

$$\overline{IC} = \frac{\sum_{j=1}^{J} \mathbf{1}(H_j < LI_{95\,\%} \text{ ó } H_j > LS_{95\,\%})}{J}$$

  El intervalo de confianza se contruye a partir de la raiz cuadrada de $\widehat{MSE}_B(\hat{\bar{H}}_j)$ y asumiendo normalidad de la distribución muestral de $\hat{\bar{H}}_j$. Esto es:

$$LI_{95\,\%} = \hat{\bar{H}}_j - z_{0.975} * \sqrt{\widehat{MSE}_B(\hat{\bar{H}}_j)} \qquad LS_{95\,\%} = \hat{\bar{H}}_j + z_{0.975} * \sqrt{\widehat{MSE}_B(\hat{\bar{H}}_j)}$$

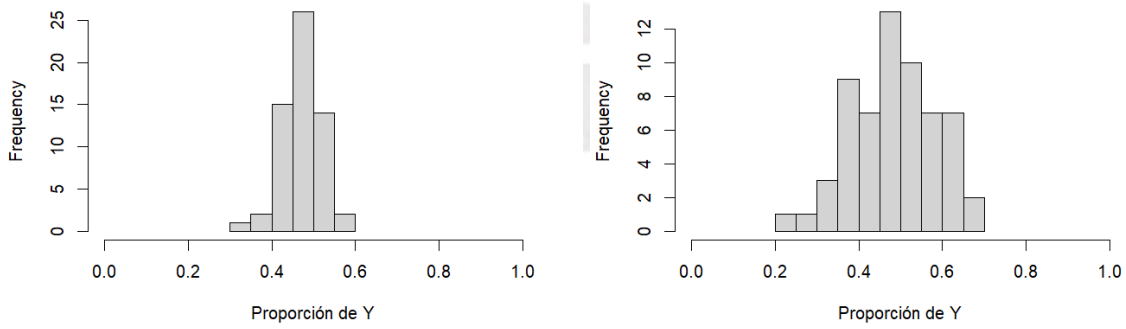
Asimismo, se realizará un análisis más desagregado para comprender el cambio de las tres estadísticas ante diferentes tamaños de muestra de las áreas.

Los escenarios simulados considerarán los siguientes aspectos generales:

- Sesenta áreas, es decir $J = 60$, con un tamaño de su población $N_j = 800$ por área que, en su conjunto, equivalen a una población de tamaño $N = 48000$.
- Las observaciones son muestreadas aleatoriamente (sin reemplazo) en cada una de las 60 áreas. Para ello, se considera diferentes tamaños de muestra que son elegidos aleatoriamente entre las posibilidades $n_j =$ 10, 20, 50 ó 100.
- Para los cálculos de $\hat{\bar{H}}_j$ y $\widehat{MSE}_B(\hat{\bar{H}}_j)$ se considerará $L = 200$ y $B = 200$ para las estimaciones por Monte Carlo y $bootstrap$ respectivamente.

4.1. Simulaciones bajo el supuesto de una variable dependiente de tipo dicotómica

El primer ejercicio de simulación consiste en la generación de datos artificiales de tipo dicotómico para una variable dependiente de interés Y, cuyo valor 1 lo asociaremos a un éxito. Su proceso generador de datos se simula con $logit(\pi_{ij})) = -1.1 + 0.1 * X_{ij} + \gamma_j$, donde $X_{ij} \sim N(10, 5)$, y $\gamma_j \sim N(0, \sigma^2_\gamma)$, donde usaremos los valores 0.02 y 0.12 para σ^2_γ. Como resultado, las proporciones poblacionales de la variable dependiente $H_j = \pi_j = \frac{1}{N_j} \sum_{i=1}^{N_j} Y_{ij}$ para las sesenta áreas se distribuyen de la siguiente manera:



(a) σ^2_γ = 0.02 **(b)** σ^2_γ = 0.12

Figura 3: Distribución de las proporciones poblacionales π_j (caso: datos dicotómicos) bajo diferentes valores de σ^2_γ.
Elaboración propia.

Tomando estos datos en consideración, se realiza el muestreo aleatorio a las sesenta

áreas para luego obtener las estimaciones directas, las estimaciones por regresión lineal de error anidado (BHF) y por GAMLSS del promedio de Y para cada área, así como las tres estadísticas indicadas anteriormente para comparar su robustez. En el caso del GAMLSS, se aplican los métodos descritos en la sección 3.2 para obtener $\hat{\bar{H}}_j = \hat{\bar{\pi}}_j$ y $\widehat{MSE}_B(\hat{\bar{H}}_j) = \widehat{MSE}_B(\hat{\bar{\pi}}_j)$ para $j = 1, \ldots, 60$, considerando una función de enlace $logit$ para $g_\mu(\mu_{ij}) = g_\mu(\pi_{ij})$ y los datos muestrales de X_{ij}. Los resultados agregados de esta simulación, según los valores de σ^2_γ se resumen en la Tabla 1.

	σ^2_γ = 0.02			σ^2_γ = 0.12		
	Est. Directa	Est. BHF	Est. GAMLSS	Est. Directa	Est. BHF	Est. GAMLSS
$\overline{MSE}$	0.0098	0.0015	0.0016	0.0098	0.0035	0.0038
$\overline{Diff}$	0.1618	0.0742	0.0749	0.1443	0.1253	0.1263
$\overline{IC}$	0.1000	0.0500	0.0667	0.0000	0.0167	0.0500

Tabla 1: Comparación de las estadísticas agregadas de $\hat{\bar{H}}_j = \hat{\bar{\pi}}_j$ bajo la simulación de datos dicotómicos, según tipo de estimación y nivel de variabilidad de efectos aleatorios de las áreas.
Elaboración propia a partir de estimaciones realizadas en R.

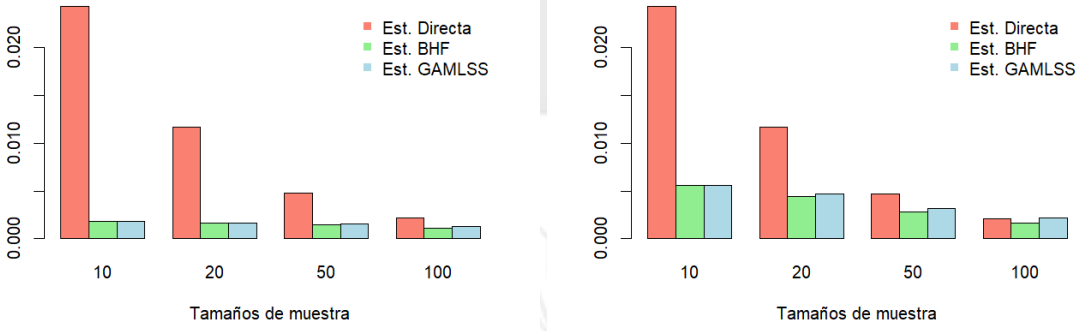
Se observa que las estimaciones por GAMLSS utilizando las muestras simuladas calculan resúmenes de Y a nivel de área con menor error cuadrático medio y menor distancia respecto a sus valores poblacionales, al compararlo con las estimaciones directas, independientemente del valor de σ^2_γ. Al compararlo con el modelo de Battese, Harter y Fuller (1988), los resultados son similares, es decir no se evidencian diferencias importantes en el error cuadrático medio ni en la distancia de los indicadores respecto a su valor poblacional.

No obstante, la mayor variabilidad de los efectos aleatorios de las áreas provoca un aumento en la varianza de las estimaciones por GAMLSS, traduciéndose en un ligero incremento de su error cuadrático medio y en una mayor distancia respecto a las medias poblacionales. Respecto a la proporción de áreas cuyo valor poblacional no se encuentra en el intervalo de confianza al 95 %, su cifra es reducida para los tres métodos de estimación (no supera el 10 % en ningún caso).

Desagregando las estadísticas según tamaños de muestra, se evidencia que la estimación por GAMLSS es relativamente más robusta que la estimación directa, especialmente bajo tamaños de muestra reducidos donde la dispersión de las estimaciones directas es elevada. Como se observa en la Figura 4, el error cuadrático medio de las estadísticas a nivel de área calculadas mediante GAMLSS es más bajo que las obtenidas de manera directa, especialmente cuando el tamaño de la muestra para las áreas es de 10 o 20, y es similar al obtenido mediante el modelo de regresión lineal de error anidado (BHF). Se evidencia una convergencia a la igualdad del error cuadrático medio de los tres métodos

a partir de $n_j = 100$, independientemente de la variabilidad de los efectos aleatorios de las áreas. Esto es consistente con del Teorema de Límite Central, que indica que la distribución muestral de la $\hat{\bar{H}}_j$ empezará a converger a una normal con varianza inversamente proporcional al tamaño de muestra.

En el caso de las diferencias en valor absoluto entre $\hat{\bar{H}}_j$ y H_j, como se observa en la Figura 5, el tamaño de muestra también influye en la exactitud de las estadísticas a nivel de área de la variable de interés. En general, a menor tamaño de muestra y mayor σ^2_γ mayor es la diferencia entre el valor estimado y el valor real de los promedios a nivel de área de Y, así como su dispersión. No obstante, las estimaciones por GAMLSS siempre generan diferencias menores o similares a las estimaciones directas, y tienden a coincidir con las obtenidas a través del modelo BHF.



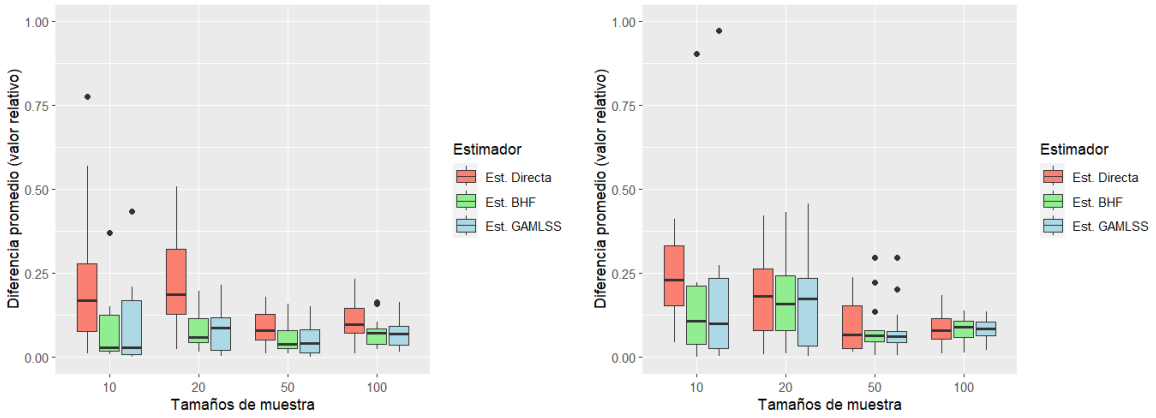
(a) σ^2_γ = 0.02 **(b)** σ^2_γ = 0.12

Figura 4: Error cuadrático medio (MSE) según tamaños de muestra de áreas y valores de σ^2_γ (caso: datos dicotómicos).
Elaboración propia a partir de estimaciones realizadas en R.

4.1.1. Caso: datos desbalanceados

La presencia de datos desbalanceados en una variable dicotómica, donde predomine fuertemente alguna de sus dos categorias, representa un obstáculo en la formulación de modelos predictivos. En este caso, las estimaciones tenderán a favorecer la categoría más prevalente en los datos, lo que puede introducir sesgos en las predicciones. Dicha situación podría ser costosa o peligrosa en caso se modelen eventos con bajas tasas de ocurrencia como el fracaso de la adquisición de una empresa o la ocurrencias de fraudes financieros (Lee, Joo, Baik, Han y In, 2020).

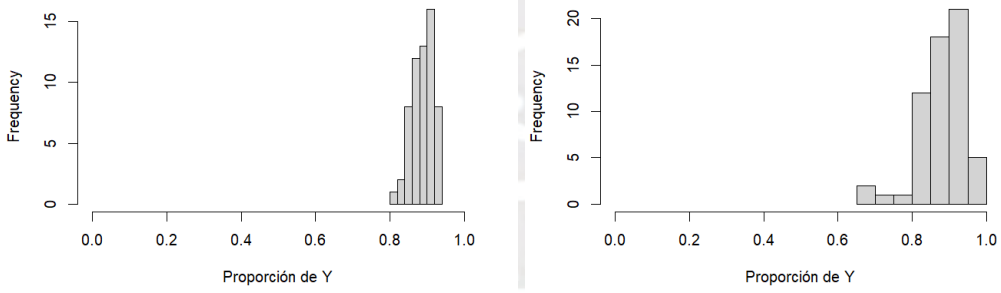
Tomando ello en consideración, resulta de interés comprobar si la estimación en áreas pequeñas con GAMLSS es resiliente a dicho fenómeno. Para ello, se simulan datos de tipo



(a) σ^2_γ = 0.02 **(b)** σ^2_γ = 0.12

Figura 5: Gráficos de caja de las diferencias promedio en valor absoluto entre $\hat{\bar{H}}_j$ y H_j según tamaños de muestra de áreas y valores de σ^2_γ (caso: datos dicotómicos).
Elaboración propia a partir de estimaciones realizadas en R.

dicotómico considerando un esquema generador de datos similar al anterior $logit(\pi_{ij}) = 1.3 + 0.08 * X_{ij} + \gamma_j$, donde $X_{ij} \sim N(10, 5)$ y $\gamma_j \sim N(0, \sigma^2_\gamma)$, generando las distribuciones de las proporciones de la variable dependiente $H_j = \pi_j$ según se observa en la Figura 6. Nótese cómo estos promedios se encuentran en su mayoría entre 0.8 y 1 especialmente en el caso donde la varianza de los efectos aleatorios de las áreas es baja, denotando una elevada frecuencia de la categoría de éxito en todas las áreas.



(a) σ^2_γ = 0.06 **(b)** σ^2_γ = 0.24

Figura 6: Distribución de las proporciones poblacionales π_j (caso: datos desbalanceados) bajo diferentes valores de σ^2_γ.
Elaboración propia.

Repitiendo el proceso de estimación realizado anteriormente, se obtienen las estadísticas agregadas para la comparación del método GAMLSS respecto a la estimuación directa considerando el diseño muestral y el modelo de Battese, Harter y Fuller (1988). Como se observa en la Tabla 2, los cálculos muestran que las estimaciones utilizando el método GAMLSS tienen menores errores cuadráticos medios y diferencias con los valores prome-

dios poblacionales en comparación con las estimaciones directas, siendo ello consistente con lo descrito en la Tabla 1.

Por otro lado, bajo el escenario de datos desbalanceados, el método de estimación directa incurre en mayores sesgos, reflejándose ello en la mayor proporción de áreas cuyo promedio real no se encuentra en el intervalo de confianza al 95 % de su valor estimado. En el caso de los GAMLSS, el valor de esta estadística es bajo y similar al obtenido bajo el supuesto de datos balanceados. Este análisis demuestra que la aplicación de los GAMLSS en un contexto de estimación en áreas pequeñas es resiliente (o al menos más resiliente que las estimaciones directas) ante la presencia de datos dicotómicos desbalanceados.

En cuanto a las semejanzas de las estadísticas agregadas del método GAMLSS con el método BHF, estas se mantienen, aún bajo el supuesto de datos desbalanceados. Finalmente, al desagregar el análisis según tamaño de muestra, no se encontraron mayores hallazgos, manteniendose el patrón encontrado en el anterior apartado.

	σ^2_γ = 0.06			σ^2_γ = 0.24		
	Est. Directa	Est. BHF	Est. GAMLSS	Est. Directa	Est. BHF	Est. GAMLSS
$\overline{MSE}$	0.0039	0.0006	0.0006	0.0042	0.0013	0.0015
$\overline{Diff}$	0.0454	0.0231	0.0232	0.0494	0.0380	0.0380
$\overline{IC}$	0.0833	0.0667	0.0500	0.1000	0.0833	0.0667

Tabla 2: Comparación de las estadísticas agregadas de $\hat{\bar{H}}_j = \hat{\bar{\pi}}_j$ bajo la simulación de datos dicotómicos no balanceados, según tipo de estimación y nivel de variabilidad de efectos aleatorios de las áreas.
Elaboración propia a partir de estimaciones realizadas en R.

4.2. Simulaciones bajo el supuesto de una variable dependiente de tipo numérica no normal

Cómo se ha mencionado en el marco teórico, la principal ventaja de los GAMLSS es la flexibilidad que ofrece en la modelización de la variable dependiente, siendo esto útil cuando nos enfrentamos a datos no normales con elevada asimetria o curtosis. En ese sentido, es de interés evaluar el rendimiento de las estimaciones por GAMLSS bajo este tipo de escenarios.

4.2.1. Caso: datos discretos

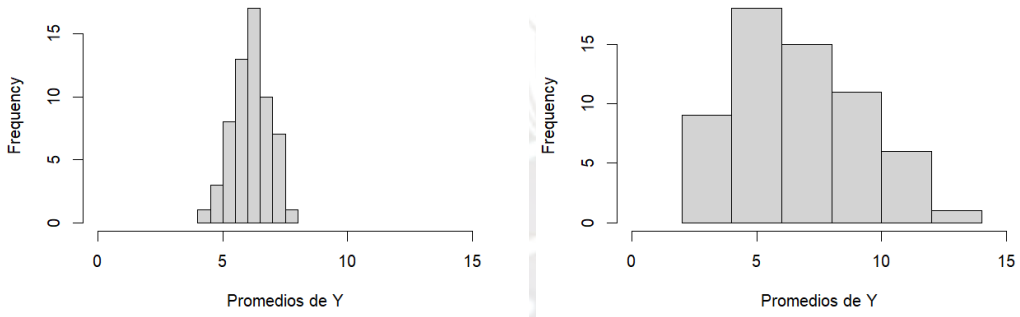
El primer caso a evaluar será el de un proceso generador de datos de tipo discreto según una distribución $Binomial\ Negativa$. Para realizar la modelización de una variable

dependiente y de interés bajo esta distribución con un GAMLSS, Rigby y otros (2019) plantean la siguiente reparametrización de su función de probabilidad:

$$f(y|\mu,\sigma) = \frac{\Gamma(y+\frac{1}{\sigma})}{\Gamma(\frac{1}{\sigma})\Gamma(y+1)} \left(\frac{\sigma\mu}{1+\sigma\mu}\right)^{y} \left(\frac{1}{1+\sigma\mu}\right)^{1/\sigma}$$

Note que bajo esta especificación, es posible realizar la modelización de tanto la ubicación (μ) como la dispersión (σ) de la distribución de datos discretos. Considerando ello, se propone un proceso generador de datos según $log(\mu_{ij}) = 1.5 + 0.03 * X_{ij} + \gamma_j$ y $log(\sigma_{ij}) = -2.2 + 0.08 * X_{ij} + \gamma_j$, donde $X_{ij} \sim N(10, 5)$ y $\gamma_j \sim N(0, \sigma^2_\gamma)$ (valores alternativos de σ^2_γ = 0.01 y 0.12).

Con ello, se simulan datos aleatorios de Y_{ij} y se realiza el muestreo correspondiente para cada una de las sesenta áreas, siguiendo los lineamientos generales indicados al inicio del presente capítulo. Como resultado, los valores promedio poblacionales de la variable dependiente $H_j = \mu_j$ para las sesenta áreas se distribuyen según lo indicando en la Figura 7.



(a) σ^2_γ = 0.01 **(b)** σ^2_γ = 0.12

Figura 7: Distribución de los promedios poblacionales μ_j (caso: datos discretos) bajo diferentes valores de σ^2_γ. Elaboración propia.

Como se puede observar, la distribución $Binomial\ Negativa$ afecta los promedios de las áreas, provocando una asimetría hacia la derecha especialmente cuando sus efectos aleatorios presentan mayor varianza. Con estos datos generados se realiza las estimaciones pertinentes para obtener las estadísticas de comparación de robustez, especificamente el error cuadrático medio, la diferencia relativa promedio respecto a los valores poblacionales y la proporción de áreas cuyo valor poblacional no se encuentre dentro del intervalo de confianza de su estimador. Las estadísticas agregadas se muestran en la Tabla 3.

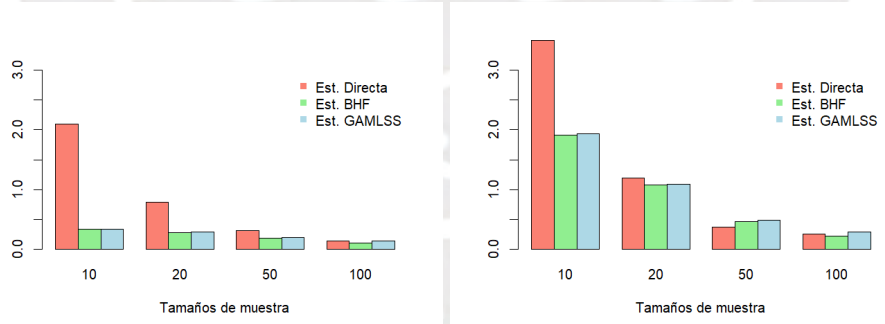
De acuerdo a estas cifras, se mantendría lo hallado en las simulaciones realizadas bajo el supuesto de una variable dependiente de tipo binaria: las estimaciones de indicadores

	σ^2_γ = 0.01			σ^2_γ = 0.12		
	Est. Directa	Est. BHF	Est. GAMLSS	Est. Directa	Est. BHF	Est. GAMLSS
$\overline{MSE}$	0.7518	0.2175	0.2346	1.1941	0.8535	0.8868
$\overline{Diff}$	0.0871	0.0682	0.0715	0.0884	0.1046	0.0985
$\overline{IC}$	0.0833	0.0667	0.0500	0.0833	0.0167	0.0167

Tabla 3: Comparación de las estadísticas agregadas de $\hat{\bar{H}}_j = \hat{\bar{\mu}}_j$ bajo la simulación de datos numéricos discretos según tipo de estimación y nivel de variabilidad de efectos aleatorios de las áreas.
Elaboración propia a partir de estimaciones realizadas en R.

en áreas pequeñas mediante GAMLSS tendrían errores cuadrático medio y diferencias relativas promedio similares a las estimaciones obtenidas mediante un modelo de Battese, Harter y Fuller (1988) y menores a las estimaciones realizadas de manera directa. Asimismo, se evidencia menores valores de la proporción de áreas cuyo valor poblacional no se encuentra dentro del intervalo de confianza de su estimador. Estos resultados son invariables bajo distintos valores de la variabilidad de los efectos aleatorios de las áreas.

Al desagregar estos resultados según tamaños de muestra, se encuentra que las bondades de los GAMLSS son más evidentes bajo tamaños de muestra reducidos especialmente al contrastar los valores de error cuadrático medio. Bajo este escenario de simulación de una variable dependiente numérica de tipo discreta, la robustez de los métodos se empieza a equiparar a partir de un tamaño de muestra de 50 unidades por área.

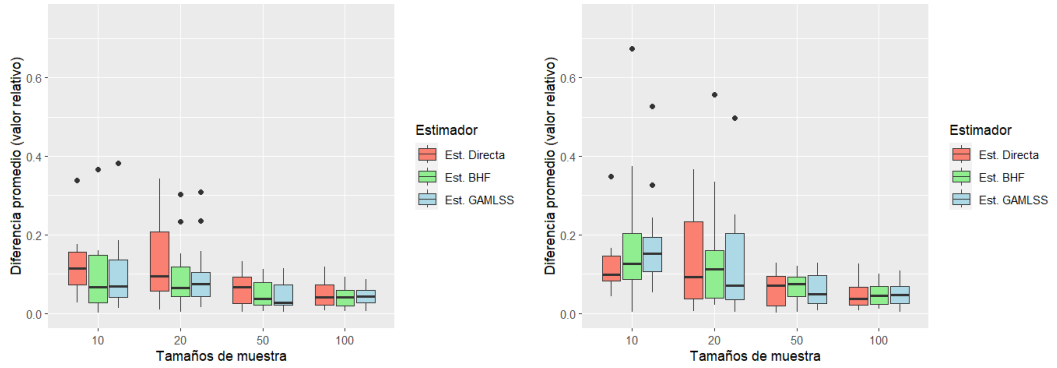


(a) σ^2_γ = 0.01 **(b)** σ^2_γ = 0.12

Figura 8: Error cuadrático medio (MSE) según tamaños de muestra de áreas y valores de σ^2_γ (caso: datos numéricos discretos).
Elaboración propia a partir de estimaciones realizadas en R.

4.2.2. Caso: datos continuos

Un último caso a evaluar en las simulaciones será el de una variable dependiente numérica continua y con distribución no normal. Para ello, se elige generar de datos artificiales



(a) σ^2_γ = 0.01 **(b)** σ^2_γ = 0.12

Figura 9: Gráficos de caja de las diferencias promedio en valor relativo entre $\hat{\bar{H}}_j$ y H_j según tamaños de muestra de áreas y valores de σ^2_γ (caso: datos numéricos discretos).
Elaboración propia a partir de estimaciones realizadas en R.

con distribución $Gamma$ considerando la siguiente reparametrización de su función de densidad, ajustada al enfoque de modelización de los GAMLSS por Rigby y otros (2019):

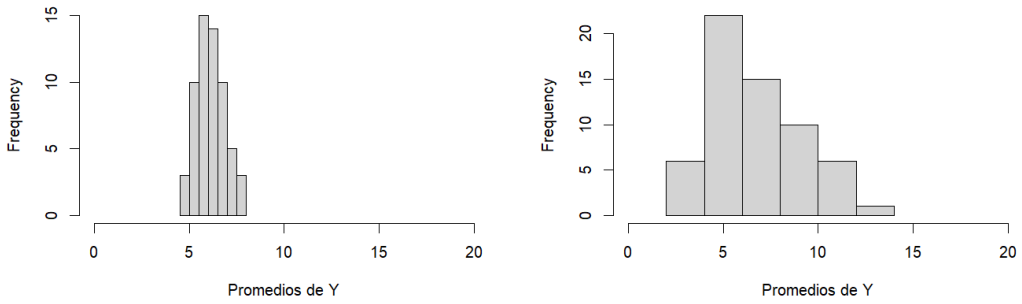
$$f(y|\mu,\sigma) = \frac{y^{(1/\sigma^2-1)}\exp[-y/(\sigma^2\mu)]}{(\sigma^2\mu)^{(1/\sigma^2)}\Gamma(1/\sigma^2)}$$

Bajo esta especificación, los parámetros de la distribución $Gamma(\alpha,\beta)$, se reparametrizan a $\alpha_{ij} = 1/\sigma^2_{ij}$ y $\beta_{ij} = 1/(\mu_{ij}\sigma^2_{ij})$, con el propósito de incorporar los parámetros de ubicación (μ) y dispersión (σ) para su modelización. Nótese además que en este caso particular, la varianza de la variable dependiente es una función de ambos parámetros, tal que $V(Y) = \frac{\alpha_{ij}}{\beta^2_{ij}} = \frac{1/\sigma^2_{ij}}{1/(\mu_{ij}\sigma^2_{ij})^2} = \frac{\mu^2_{ij}(\sigma^2_{ij})^2}{\sigma^2_{ij}} = \mu^2_{ij}\sigma^2_{ij}$. Con ello en consideración, el proceso generador de datos para los parámetros de esta distribución se plantea como $log(\mu_{ij}) = 1.5 + 0.03 * X_{ij} + \gamma_j$ y $log(\sigma_{ij}) = -0.9 + 0.01 * X_{ij} + \gamma_j$, donde $X_{ij} \sim N(10,5)$ y $\gamma_j \sim N(0,\sigma^2_\gamma)$ (valores alternativos de σ^2_γ = 0.01 y 0.1).

Repitiendo el ejercicio realizado para el caso de datos discretos, se simula la distribución de Y_{ij} con μ_{ij} y σ_{ij}. Posteriormente, se realiza el muestreo correspondiente para cada una de las sesenta áreas. Con ello se obtienen los valores promedio poblacionales de la variable dependiente $H_j = \mu_j$ para las sesenta áreas, las cuales se distribuyen según la Figura 10.

Al igual que los datos generados con la distribución $Binomial\ Negativa$, se evidencia cómo la asimetría de la distribución $Gamma$ afecta los promedios de las áreas, presentándose una mayor dispersión provocando cuando sus efectos aleatorios presentan mayor varianza.

A partir de los datos simulados, se obtienen las estadísticas agregadas para comparar el rendimiento de los GAMLSS con la estimación directa por diseño muestral y el modelo



(a) σ^2_γ = 0.01 **(b)** σ^2_γ = 0.1

Figura 10: Distribución de los promedios poblacionales μ_j (caso: datos numéricos continuos) bajo diferentes valores de σ^2_γ.
Elaboración propia.

de Battese, Harter y Fuller (1988) (ver Tabla 4). En general, los resultados son consistentes con lo encontrado en los casos anteriores: los indicadores a nivel de área $\hat{\bar{H}}_j$ estimados por GAMLSS tienen en promedio un menor error cuadrático medio que las estimacioes directas, independientemente de la variabilidad de los efectos aleatorios de las áreas.

En el caso de la comparación con las estimaciones de la regresión lineal de error anidado de Battese, Harter y Fuller (1988), los errores cuadráticos medios son similares excepto ante la presencia de mayor varianza de γ_j, donde el error cuadrático medio del modelo BHF es mayor. Esto se debería a la mayor asimetría de los datos que provoca el incumplimiento del supuesto de normalidad de errores del modelo de BHF.

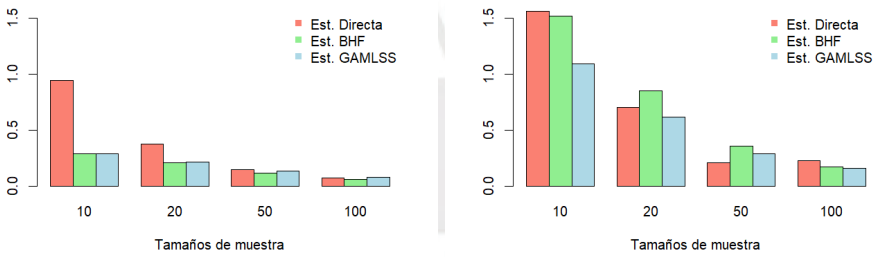
	σ^2_γ = 0.01			σ^2_γ = 0.1		
	Est. Directa	Est. BHF	Est. GAMLSS	Est. Directa	Est. BHF	Est. GAMLSS
$\overline{MSE}$	0.3512	0.1629	0.1722	0.6264	0.6740	0.5044
$\overline{Diff}$	0.0586	0.0498	0.0474	0.0758	0.0751	0.0732
$\overline{IC}$	0.0000	0.0333	0.0167	0.0833	0.0500	0.0833

Tabla 4: Comparación de las estadísticas agregadas de $\hat{\bar{H}}_j = \hat{\bar{\mu}}_j$ bajo la simulación de datos numéricos continuos según tipo de estimación y nivel de variabilidad de efectos aleatorios de las áreas.
Elaboración propia a partir de estimaciones realizadas en R.

Respecto a la brecha de los indicadores estimados con sus valores poblacionales, las estimaciones por GAMLSS muestran menores sesgos que las estimaciones directas y bastante similitud con los resultados del modelo BHF, si los efectos aleatorios de las áreas no son muy variables. Sin embargo, bajo el escenario de σ^2_γ = 0.1, no se evidencian diferencias importantes en el valor promedio de la diferencia relativa entre $\hat{\bar{H}}_j$ y $\bar{H}_j$ ni en la proporción de áreas cuyo valor de $\bar{H}_j$ no se encuentre en el intervalo de confianza al 95 % de $\hat{\bar{H}}_j$, entre

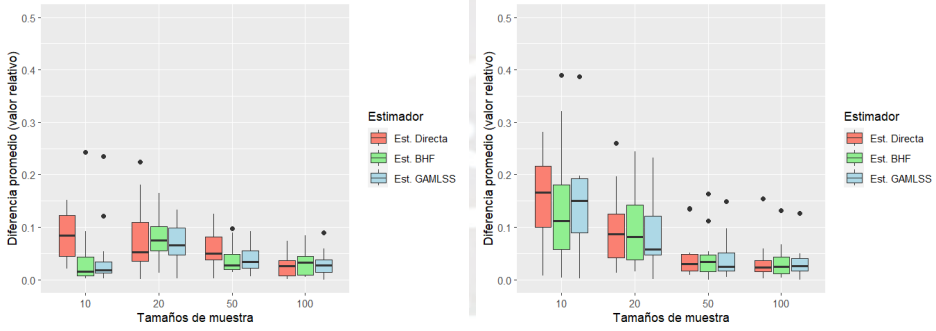
ninguno de los tres métodos de estimación.

Para estudiar este fenómeno en mayor detalle, se realiza un análisis desagregando las estadísticas según tamaño de muestra. Como se puede observar en las Figuras 11 y 12, bajo el escenario de baja heterogeneidad, las estadísticas se comportan de la manera esperada. En el caso de las estimaciones por GAMLSS, mientras mayor tamaño tengan las muestras de las áreas, menor es su error cuadrático medio, así como su diferencia promedio con respecto al valor real de H_j. Estas cifras son siempre menores a las obtenidas mediante estimaciones directas (especialmente si el tamaño de muestra es bajo) y muy similares a las del modelo BHF.



(a) $\sigma^2_\gamma = 0.01$ **(b)** $\sigma^2_\gamma = 0.1$

Figura 11: Error cuadrático medio (MSE) según tamaños de muestra de áreas y valores de σ^2_γ (caso: datos numéricos continuos).
Elaboración propia a partir de estimaciones realizadas en R.



(a) $\sigma^2_\gamma = 0.01$ **(b)** $\sigma^2_\gamma = 0.1$

Figura 12: Gráficos de caja de las diferencias promedio en valor relativo entre $\hat{\bar{H}}_j$ y H_j según tamaños de muestra de áreas y valores de σ^2_γ (caso: datos numéricos continuos).
Elaboración propia a partir de estimaciones realizadas en R.

Sin embargo, bajo el escenario de elevada heterogeneidad entre las áreas, se evidencia un aumento generalizado del error cuadrático medio y de la diferencia relativa entre el valor poblacional y el valor estimado de las características, especialmente cuando los tamaños de muestra son reducidos. A pesar de ello, la estimación por GAMLSS mantiene el menor error cuadrático medio de los tres métodos de estimación. En el caso de las diferencias

entre $\hat{\tilde{H}}_j$ y H_j se evidencia nuevamente que el método GAMLSS genera la menor brecha, aunque solo cuando los tamaños de muestra son más reducidos.

Capítulo 5
Aplicación de GAMLSS para obtener indicadores de infraestructura de hogares a nivel distrital en Perú

En vista de las propiedades de las estimaciones mediante GAMLSS en un contexto de estimación en áreas pequeñas, en el presente capítulo se aplicará dicho método a datos reales de Perú con el propósito de aportar a la generación de información de calidad a nivel distrital. Esta aplicación es de particular interés para la toma de decisiones de las autoridades locales, quienes cuentan con datos limitados sobre sus territorios.

La base de datos que se utilizará es la Encuesta Nacional de Hogares correspondiente al año 2017, acotada a las viviendas particulares (casas independientes, departamentos en edificios, viviendas en quinta y viviendas en casa de vecindad) del departamento de Ica[9]. De acuerdo a su documentación, el muestreo de la encuesta nacional es aleatorio, estratificado y multietápico. Es decir, que la selección se realiza por etapas: primero se elige el centro poblado (Unidad Primaria de Muestro); en segundo lugar, el conglomerado (Unidad Secundaria de Muestro); y finalmente, la vivienda (Unidad Terciaria de Muestreo).

La elección del año 2017 se debe a que durante dicho periodo se realizó un censo nacional, en donde se recogió información para obtener los valores poblacionales de indicadores vinculados a variables socioeconómicas como la condición de las viviendas, la ubicación geográfica y el acceso a servicios básicos, datos que también se recogen en la encuesta nacional pero a través del diseño muestral descrito anteriormente. Ello posibilita la comparación de las estimaciones realizadas a partir de las muestras recogidas en la encuesta con los indicadores recogidos en el censo.

En esta aplicación, se abordarán dos variables de interés:

- El acceso a servicios de saneamiento conectados a una red pública de alcantarillado dentro de los hogares. Esto corresponde a una variable de tipo dicotómica, cuyo indicador a nivel de área (en este caso, distrito) es la proporción de viviendas que tienen servicios higiénicos conectados a una red pública de desagüe dentro de la vivienda.
- La cantidad de habitaciones de las viviendas (sin contar baño, cocina, pasadizos ni garaje) que corresponde a una variable de tipo numérica discreta, y cuyo indicador a nivel de área es la cantidad promedio de habitaciones.

De acuerdo a los datos del censo nacional del año 2017, el departamento de Ica tiene

[9]Se eligió este departamento por tener un número no muy elevado de distritos con tamaños de muestra suficientemente variados para evaluar la robustez del método bajo diferentes valores de n_j.

43 distritos, cinco provincias y un total de 280 303 viviendas[10]. Por su lado, la Encuesta Nacional de Hogares del 2017 tiene información para 37 de los 43 distritos de Ica, así como para sus cinco provincias. Revisando la Figura 13, se puede observar que existe una elevada variabilidad en algunos de los indicadores poblacionales a nivel de área. Se tienen distritos con proporciones de viviendas con acceso a saneamiento por debajo de 0.2 mientras que otras áreas tienen una cobertura de saneamiento mejorado por encima de 0.8, siendo el valor promedio 0.62. En el caso de la cantidad promedio de habitaciones, se tiene menor variabilidad, siendo el valor mínimo 1.4 y el valor máximo 3.4.

Respecto a los tamaños de muestra de los distritos coberturados en la Encuesta Nacional de Hogares, se observa una elevada dispersión: se tienen áreas con tamaños de muestra menores a 10 viviendas mientras que otros distritos cuentan con muestras mayores a 100. A pesar de ello, la mayor cantidad de áreas cuentan con tamaños de muestra entre 16 y 47 viviendas, siendo su valor mediano igual a 22 y la sumatoria de todas las áreas igual a 1412.

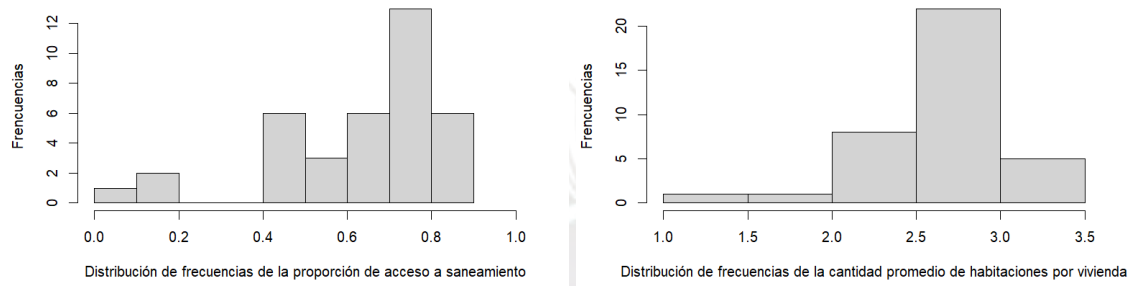


Figura 13: Distribución de indicadores poblacionales (H_j) de las variables de interés de los distritos de Ica coberturados en la encuesta nacional para los distritos $j = 1, \ldots, 37$.
Elaboración propia a partir de datos del Censo Nacional de Población y Vivienda 2017

5.1. Estimación GAMLSS de indicadores de infraestructura para los distritos del departamento de Ica

Tomando en consideración estos datos, se aplica la metodología GAMLSS para realizar la estimación de los indicadores de acceso a saneamiento por red pública dentro de las viviendas y cantidad de habitaciones de las viviendas para cada uno de los 37 distritos de Ica contenidos en la muestra de la ENAHO. Si bien se sabe que dicha encuesta no tiene sufiente información para realizar inferencia a nivel provincial o distrital, se espera que las

[10] Información obtenida del Sistema de Consulta de Base de Datos - REDATAM del Censo Nacional 2017 del INEI.

estimaciones por GAMLSS tengan menor varianza y que sean más precisas, especialmente en aquellos distritos con menor tamaño de muestra. A continuación, se muestran los principales resultados según cada indicador.

5.1.1. Proporción de viviendas que tienen servicios higiénicos conectados a una red pública de desagüe dentro de la vivienda

De acuerdo a los datos de la ENAHO para el departamento de Ica, casi 8 de cada 10 viviendas particulares muestreadas tienen acceso a servicios higiénicos conectados a una red pública de desagüe dentro de la vivienda. Realizando un cruce con otras variables disponibles en la encuesta (ver Figura 14), se halla que el acceso a saneamiento es más común en viviendas ubicadas en el ámbito urbano y que cuentan con un material predominante de piso diferente a tierra (parquet, lozetas, madera, cemento, etc.). Asimismo, se evidencia mayor acceso en viviendas cuyos hogares corresponden a un estrato socioeconómico alto y con algún miembro de familia que cuente con nivel de educación superior. Finalmente, se encuentra que los hogares que residen en viviendas con acceso a servicios higiénicos conectados a una red pública gastan en promedio más por su consumo de agua que aquellos hogares que no cuentan con acceso a dicho servicio.

Tomando estos resultados en consideración, para la modelización de esta variable dicotómica, se considerará una función de enlace $logit$, así como las siguientes covariables contenidas en la Encuesta Nacional de Hogares para estimar el parámetro de su distribución $Bernoulli$ mediante GAMLSS:

- El ámbito geográfico donde se ubica la vivienda: 1 = ámbito urbano, 0 = ámbito rural (variable dicotómica).
- El material de las paredes de la vivienda: 1 = cemento, 0 = otro material (variable dicotómica).
- El material del piso de la vivienda: 1 = piso de tierra, 0 = otro material (variable dicotómica).
- La pertenencia de la vivienda: 1 = la vivienda le pertenece al jefe del hogar, 0 = la vivienda no le pertenece al jefe del hogar (variable dicotómica).
- El gasto por el servicio del agua dentro de la vivienda (variable numérica).
- El mayor nivel educativo de los miembros del hogar: 1 = si algún miembro del hogar ha alcanzado un nivel educativo superior, 0 = de otro modo (variable dicotómica).

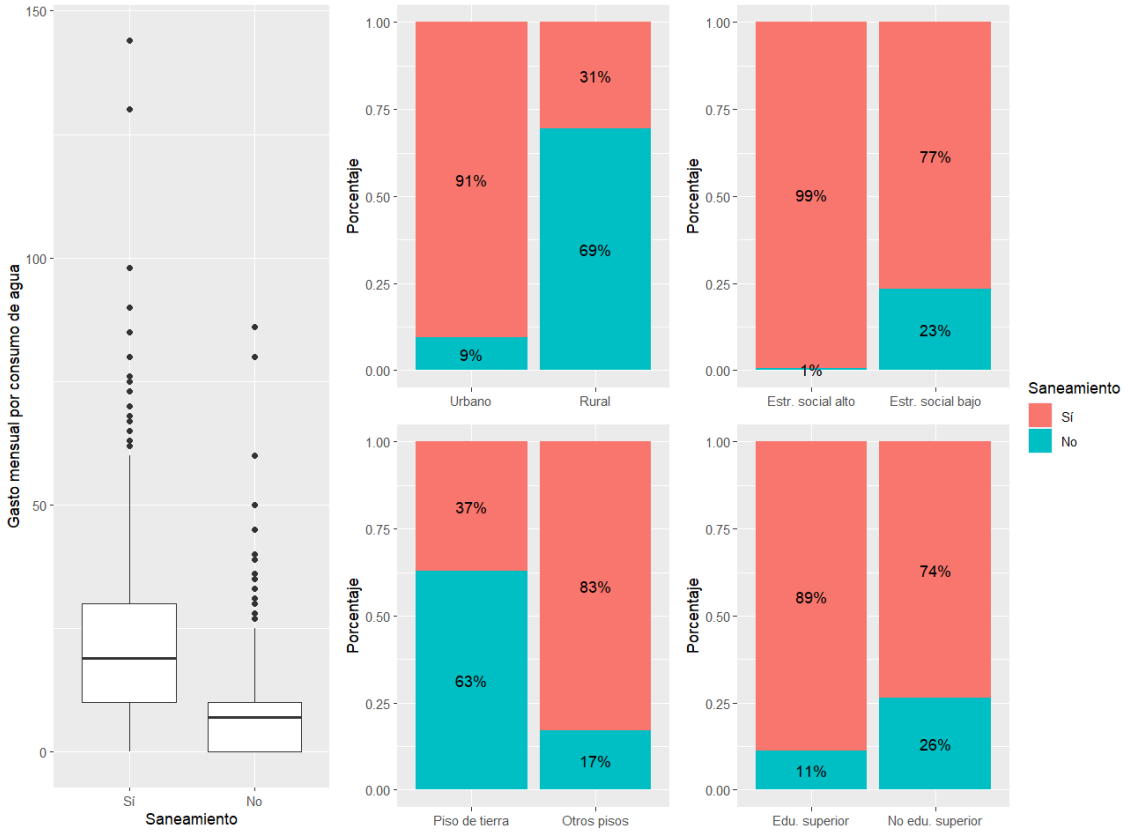


Figura 14: Análisis descriptivo del acceso a servicios higiénicos conectados a una red pública de desagüe dentro de la vivienda.
Elaboración propia a partir de la ENAHO 2017.

- La percepción del jefe del hogar respecto a los gobiernos locales: 1 = muy mala, 0 = resto de percepciones (variable dicotómica).
- El estrato socioeconómico de la vivienda: 1 = A - C, 0 = resto de estratos (variable dicotómica).
- El monto de alquiler mensual pagado o que el dueño de la vivienda cree que podría cobrar (variable numérica).
- La provincia en la que se ubica la vivienda (conjunto de cinco variables dicotómicas).

Los resultados con la ENAHO del modelo planteado en (6) para abordar el parámetro de la distribución $Bernoulli$ del acceso a saneamiento mejorado se muestran en la Figura 15 y la Tabla 5. Cabe resaltar que para ello, se han considerado como pesos[11] los factores de expansión de la encuesta que representan el inverso de su probabilidad final de selección. De esta manera, es posible incorporar el diseño muestral dentro de la estimación del

[11] Se ha verificado que la incorporación de los pesos en la estimación de los modelos es relevante, ya que su omisión genera mayores brechas entre H_j y $\hat{\hat{H}}_j$.

GAMLSS, según lo indicado en el capítulo 2.

En cuanto a la bondad de ajuste, la Figura 15 muestra que los residuos[12] del modelo planteado siguen un comportamiento aproximadamente normal como lo muestra su histograma y el Q-Q Plot. Asimismo, las pruebas de normalidad de Shapiro-Wilk y Kolmogorov-Smirnov, no rechazan la hipótesis nula de normalidad de los residuos, siendo sus *p-values* de 0.1093 y 0.2849 respectivamente.

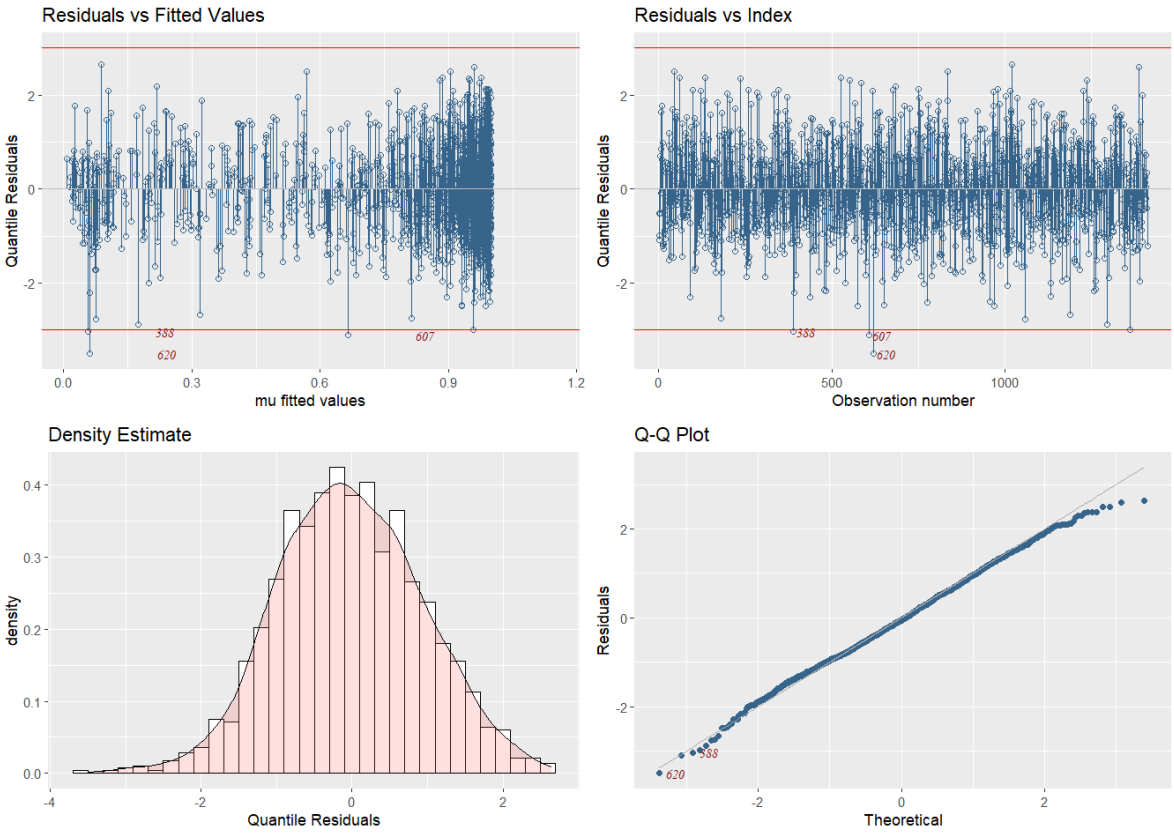


Figura 15: Análisis de residuos del modelo de acceso a servicios higiénicos conectados a una red pública de desagüe dentro de la vivienda.
Elaboración propia a partir de la ENAHO 2017.

En cuanto a los coeficientes obtenidos, sus valores se alinean a lo hallado preliminarmente en el análisis descriptivo: las covariables de ubicación geográfica en ámbito urbano, estrato socioeconómico alto, gasto por consumo de agua, nivel educativo superior y monto de alquiler tendría una relación positiva con la probabilidad que una vivienda en Ica tenga acceso a servicios higiénicos conectados a una red pública de desagüe dentro de la vivienda.

Por el contrario, si el material del piso de la vivienda es de tierra y el jefe de hogar tiene una mala percepción respecto a los gobiernos locales, entonces la probabilidad de

[12]Los residuos analizados corresponden a los residuos aleatorios cuantílicos de Dunn y Smyth (1996).

	Coeficiente	Error estándar
Intercepto	-3.178***	0.028
Ámbito geográfico - urbano	3.019***	0.022
Material de paredes - cemento	0.44***	0.018
Material de piso - tierra	-0.786***	0.026
Vivienda propia	0.37***	0.017
Gasto por consumo de agua	0.062***	0.001
Mayor nivel educativo - superior	0.466***	0.019
Percepción gobiernos locales - muy mala	-0.371***	0.016
Estrato social - A, B o C	0.881***	0.062
Alquiler	0.004***	0.000
Provincia - 1102	-0.072**	0.024
Provincia - 1103	0.59***	0.030
Provincia - 1104	1.926***	0.048
Provincia - 1105	0.092***	0.020
Global Deviance: 102 917.8		
AIC: 102 997.8		

Significancia estadística: *** 0.001 , ** 0.01 , * 0.05 , . 0.1

Tabla 5: Coeficientes del modelo GAMLSS para la estimación de μ_{ij} de la distribución $Bernoulli$ para la variable dependiente de acceso a servicios higiénicos conectados a una red pública de desagüe dentro de la vivienda.
Elaboración propia a partir de estimaciones realizadas en R. Se han utilizado los factores de expansión de la ENAHO como pesos, con el propósito de considerar el diseño muestral dentro de la estimación del modelo.

tener acceso a este servicio mejorado es menor. Finalmente, las variables asociadas al material de las paredes de la vivienda y la pertenencia de la vivienda, no tuvieron efectos estadísticos significativos a un nivel de 5 %.

Utilizado los coeficientes descritos y siguiendo el proceso metodológico descrito en el capítulo 3, se obtienen las estimaciones por GAMLSS ($\hat{\tilde{H}}_j$) de los indicadores que denotan la proporción de hogares con servicios higiénico conectados a una red pública de desagüe dentro de la vivienda, así como sus intervalos de confianza al 95 % para 37 distritos del departamento de Ica, los cuales se pueden observar en la Tabla 6. De igual manera, se han incorporado las estimaciones para las cinco provincias y para todo el departamento[13]. El contraste de estas estimaciones con sus valores poblacionales obtenidos del censo nacional se ilustran en la Figura 16.

A nivel distrital, se pueden observar diversas amplitudes de los intervalos de confian-

[13]Las estimaciones a nivel provincial y departamental se obtienen aplicando la misma metodología, solo siendo necesario adecuar el nivel de agregación de los datos generados al momento de estimar $\hat{\tilde{H}}_j$.

za dependiendo del tamaño de muestra del área. Por ejemplo, el área correspondiente al ubigeo 110101 tiene la muestra más grande correspondiente a 289 viviendas. Con ello, se ha obtenido una proporción estimada de 0.8883 con un intervalo de confianza de 0.8575 y 0.9191. En contraste, el área correspondiente al ubigeo 110113 solo tiene una muestra de 6 viviendas, con los cuales se ha obtenido una estimación de 0.7121 y un intervalo de confianza de 0.4571 y 0.9672.

N°	Ubigeo	n_j	$\hat{\bar{H}}_j$	IC (95 %)	N°	Ubigeo	n_j	$\hat{\bar{H}}_j$	IC (95 %)
Departamento									
1	11	1412	0.8145	[0.7989 - 0.8301]					
Provincias									
2	1101	713	0.8113	[0.7873 - 0.8354]	5	1104	37	0.7589	[0.6363 - 0.8816]
3	1102	314	0.8544	[0.8236 - 0.8853]	6	1105	221	0.8109	[0.7696 - 0.8522]
4	1103	127	0.7261	[0.6756 - 0.7766]					
Distritos									
7	110101	289	0.8883	[0.8575 - 0.9191]	26	110207	71	0.9794	[0.9199 - 1]
8	110102	68	0.9218	[0.8505 - 0.9931]	27	110209	5	0.0289	[0 - 0.1401]
9	110103	21	0.4703	[0.3344 - 0.6063]	28	110210	28	0.7051	[0.603 - 0.8071]
10	110104	14	0.0755	[0 - 0.266]	29	110211	6	0.9142	[0.8306 - 0.9978]
11	110105	5	0.8249	[0.614 - 1]	30	110301	50	0.6705	[0.585 - 0.756]
12	110106	100	0.9804	[0.921 - 1]	31	110303	15	0.8166	[0.5987 - 1]
13	110107	22	0.2835	[0.1219 - 0.4451]	32	110304	24	0.8721	[0.7903 - 0.9538]
14	110108	44	0.8633	[0.7679 - 0.9587]	33	110305	38	0.6709	[0.5861 - 0.7556]
15	110109	19	0.7883	[0.6182 - 0.9583]	34	110401	19	0.8201	[0.6844 - 0.9557]
16	110110	20	0.7696	[0.6357 - 0.9036]	35	110402	7	0.5605	[0.2765 - 0.8446]
17	110111	53	0.6305	[0.5365 - 0.7246]	36	110405	11	0.3554	[0.138 - 0.5728]
18	110112	38	0.7684	[0.6643 - 0.8725]	37	110501	75	0.8623	[0.8005 - 0.9241]
19	110113	6	0.7121	[0.4571 - 0.9672]	38	110503	19	0.4114	[0.2599 - 0.5629]
20	110114	14	0.0408	[0 - 0.1805]	39	110504	46	0.3287	[0.229 - 0.4285]
21	110201	97	0.9761	[0.9281 - 1]	40	110505	16	0.6795	[0.5029 - 0.8561]
22	110202	6	0.9858	[0.9005 - 1]	41	110506	13	0.9688	[0.8377 - 1]
23	110204	19	0.6562	[0.466 - 0.8463]	42	110507	28	0.8949	[0.8031 - 0.9866]
24	110205	35	0.3857	[0.2517 - 0.5197]	43	110508	24	0.9460	[0.8474 - 1]
25	110206	47	0.7717	[0.6883 - 0.8552]					

Tabla 6: Estimaciones a nivel departamental, provincial y distrital e intervalos de confianza para el indicador de proporción de viviendas que tienen servicios higiénicos conectados a una red pública de desagüe dentro de la vivienda.
Elaboración propia a partir de estimaciones realizadas en R. Debido a la escala del indicador, los límites inferiores y superiores de los intervalos de confianza se han acotado a 0 y 1 respectivamente.

No obstante, como se observa en la Figura 16, resulta evidente que en el caso de diversas áreas su valor poblacional no se encuentra dentro del intervalo de confianza al 95 % de $\hat{\bar{H}}_j$. De hecho, se identifican áreas con tamaños de muestra elevados cuyo intervalo

de confianza al 95 % no incluye a H_j. Los ubigeos 110101, 110106 y 110201 con tamaños de muestra superiores o cercanos 100 presentan intervalos de confianza más estrechos pero que fallan en contener sus proporciones poblacionales. Situación similar ocurre en la mayoría de las estimaciones a nivel provincial y departamental. El análisis más a fondo de este fenómeno se realiza en la sección 5.2.

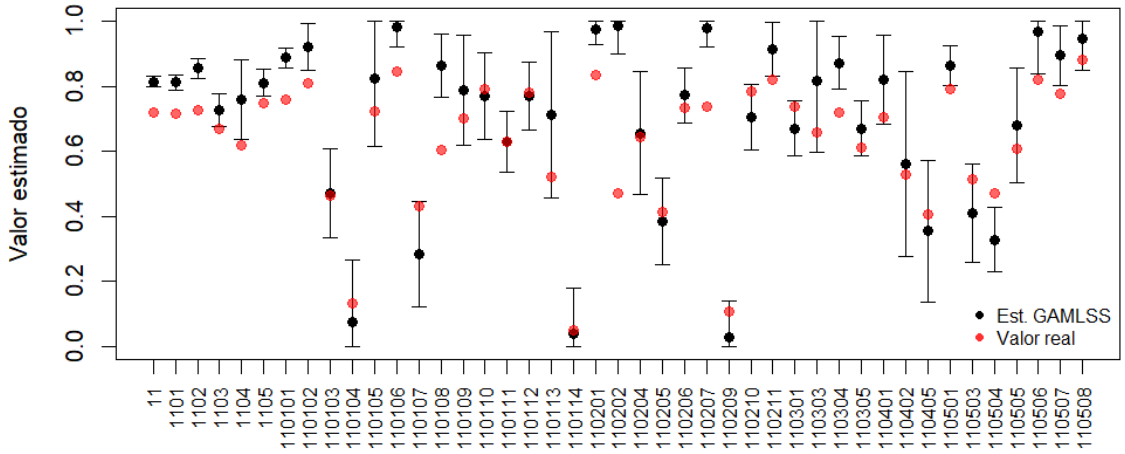


Figura 16: Intervalos de confianza al 95 % y contraste con valores poblacionales H_j del indicador de proporción de viviendas que tienen servicios higiénicos conectados a una red pública de desagüe dentro de la vivienda según ubigeos.
Elaboración propia a partir de estimaciones realizadas en R.

5.1.2. Cantidad promedio de habitaciones de la viviendas

La muestra obtenida en la ENAHO indica que una vivienda del departamento de Ica tiene en promedio 3.4 habitaciones (sin contar baño, cocina, pasadizos ni garaje). Asimismo, 26 % de las viviendas tiene dos habitaciones o menos, 55 % tienen entre tres y cuatro habitaciones, y 19 % tienen más de cuatro. Realizando un cruce con otras variables disponibles en la encuesta (ver Figura 17), se encuentra que el número de miembros del hogar, el monto de alquiler mensual y el gasto total pagado para consumo de agua, electricidad e internet son mayores en aquellas viviendas con mayor número de habitaciones. Por otro lado, los datos indican que las viviendas de estrato socioeconómico alto y donde los residentes son dueños de la vivienda, tienden a tener un mayor número de habitaciones.

En ese sentido, para la modelización[14] de esta variable numérica discreta, se considera-

[14]Antes de estimar el GAMLSS, se realizó un análisis de identificación de datos extremos utilizando las Distancias de *Cook* obtenidas a partir de un modelo lineal generalizado, siguiendo a Stevens (1984). Ello

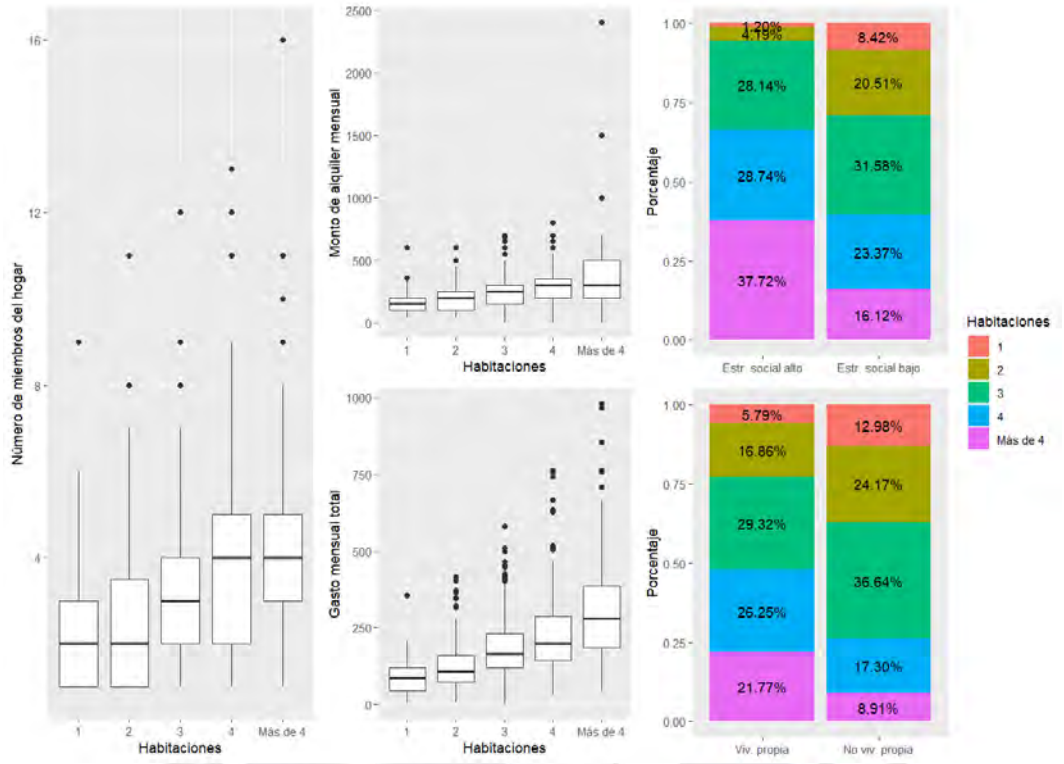


Figura 17: Análisis descriptivo de la cantidad de habitaciones dentro de las viviendas.
Elaboración propia a partir de la ENAHO 2017.

rán las siguientes covariables contenidas en la Encuesta Nacional de Hogares para estimar el parámetro de ubicación de la distribución *Binomial Negativa*[15] por GAMLSS, considerando una función de enlace logarítmica.

- El monto de alquiler mensual pagado o que el dueño de la vivienda cree que podría cobrar (variable numérica).
- El gasto total mensual por consumo de agua, electricidad, internet, entre otros gasto asociados a la vivienda (variable numérica).
- La interacción entre el alquiler mensual y el gasto total por consumo de agua, electricidad, internet, entre otros gastos (variable numérica).
- El número de miembros del hogar que residen en la vivienda (variable numérica)
- La pertenencia de la vivienda: 1 = la vivienda le pertenece al jefe del hogar, 0 = la vivienda no le pertenece al jefe del hogar (variable dicotómica).

resultó en la omisión de doce observaciones.

[15]Se elige esta distribución en lugar de una *Poisson* para aprovechar la posibilidad de modelar la media y la varianza de la distribución.

- El material del piso de la vivienda: 1 = piso de tierra, 0 = otro material (variable dicotómica).
- El mayor nivel educativo de los miembros del hogar: 1 = si algún miembro del hogar ha alcanzado un nivel educativo superior, 0 = de otro modo (variable dicotómica).
- El estrato socioeconómico de la vivienda: 1 = A - C, 0 = resto de estratos (variable dicotómica).
- La percepción del jefe del hogar respecto a los gobiernos locales: 1 = muy mala, 0 = resto de percepciones (variable dicotómica).
- La provincia en la que se ubica la vivienda (conjunto de cinco variables dicotómicas).

Asimismo, se consideran las siguientes tres covariables para modelar el parámetro de dispersión[16]. Al igual que el otro parámetro, se considera una función de enlace logarítmica para su modelización.

- El gasto total mensual por consumo de agua, electricidad, internet, entre otros gasto asociados a la vivienda (variable numérica).
- El monto de alquiler mensual pagado o que el dueño de la vivienda cree que podría cobrar (variable numérica).
- El estrato socioeconómico de la vivienda: 1 = A - C, 0 = resto de estratos (variable dicotómica).

De manera similar al modelo de acceso a saneamiento mejorado, se realiza el análisis de la bondad de ajuste del modelo de cantidad de habitaciones de las viviendas a partir del diagnóstico de sus residuos aleatorios cuantílicos. Como se muestra en la Figura 18 los residuos obtenidos se centran efectivamente alrededor de cero, sin embargo, no muestran un patrón aproximadamente normal. De acuerdo al Q-Q Plot, su distribución tendría una curtosis más positiva (colas más pesadas), lo cual es confirmado por las pruebas de normalidad de Shapiro-Wilk y Kolmogorov-Smirnov que rechazan la hipótesis nula de normalidad (*p-values* de 0.009 y 0.000 respectivamente).

Siguiendo lo indicado por Dunn y Smyth (1996), esto podría ser un indicio que los parámetros de la distribución de la variable dependiente no están siendo estimados adecuadamente. Ello podría deberse a la necesidad de incorporar otras covariables o de asumir

[16]Cabe resaltar que según Rigby y otros (2019) para valores de σ cercanos a cero, la distribución *Binomial Negativa* convergerá a una *Poisson*.

otra distribución para el modelamiento de la variable discreta. Esta situación tiene consecuencias en eficiencia de los indicadores, reflejándose ello en estimaciones mayores de la varianza de los efectos aleatorios de las áreas $\hat{\sigma}^2$.

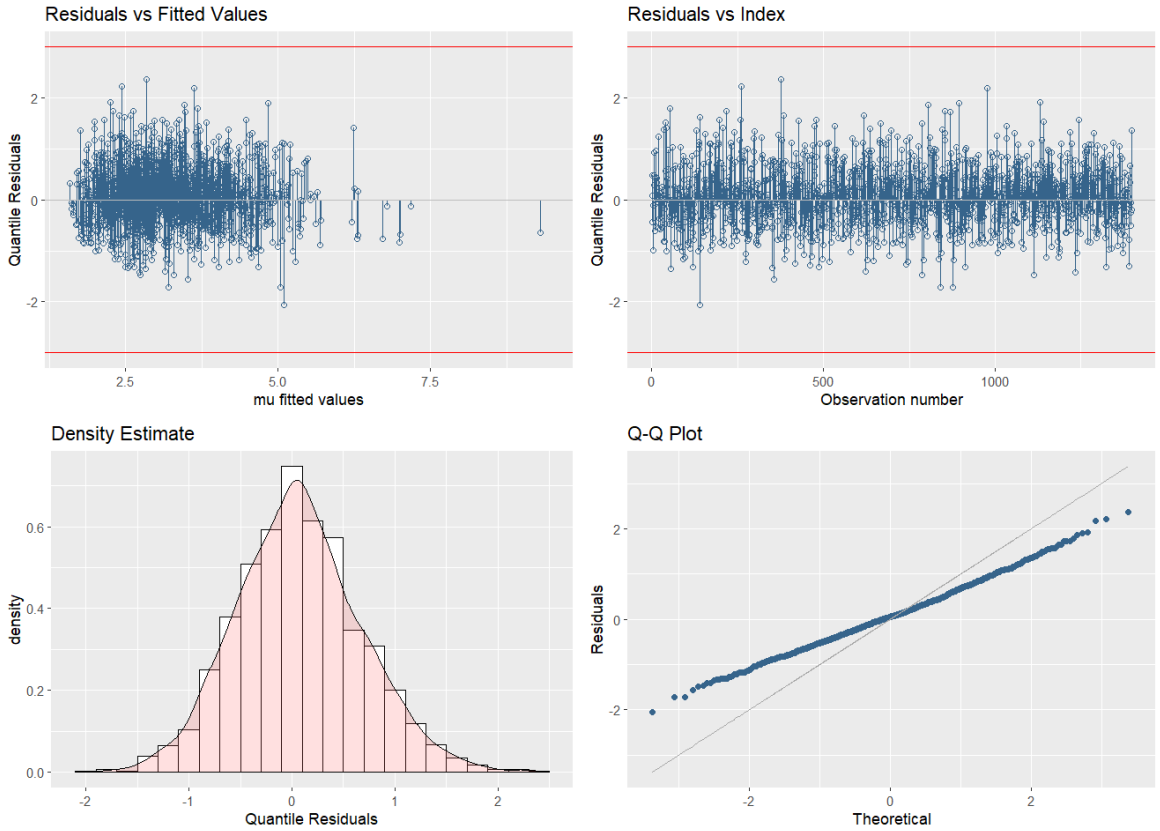


Figura 18: Análisis de residuos del modelo del número de habitaciones de las viviendas.
Elaboración propia a partir de la ENAHO 2017.

En cuanto a los coeficientes obtenidos, estos se detallan en la Tabla 7. Se encuentra que las covariables de alquiler, gasto pagado total, número de miembros del hogar y vivienda propia tendrían una relación positiva con la cantidad habitaciones dentro de viviendas. Lo contrario se encuentra para aquellas viviendas con piso de tierra y donde el jefe del hogar tiene una mala percepción de los gobiernos locales. Esto es consistente con lo hallado anteriormente en el análisis descriptivo, a excepción de la relación entre el número de habitaciones y el estrato socioeconómico, cuyo coeficiente no resultó estadísticamente significativo. Respecto a los resultados para el parámetro de dispersión, se halla que las dos primeras covariables escogidas tendrían una relación inversa con la variabilidad de la cantidad de habitaciones, mientras que la variable de estrato socioeconómico tendría una relación positiva.

De manera similar al indicador de acceso a saneamiento, las estimaciones por GAMLSS del indicador de cantidad promedio de habitaciones de las viviendas según áreas ($\hat{\tilde{H}}_j$) y sus intervalos de confianza al 95 % se muestran en la Tabla 8. Asimismo, se muestra el contras-

te de estas estimaciones con sus valores poblacionales obtenidos del censo nacional en la Figura 19.

	Coeficiente	Error estándar
Parámetro: μ		
Intercepto	0.3142***	0.00623
Alquiler	0.0016***	0.00002
Gasto pagado	0.0021***	0.00002
Alquiler x Gasto pagado	0***	0.00000
Número de miembros del hogar	0.0418***	0.00068
Vivienda propia	0.1079***	0.00302
Material de piso - tierra	-0.0465***	0.00562
Mayor nivel educativo - superior	-0.0192***	0.00278
Estrato social - A, B o C	-0.001	0.00502
Percepción gobiernos locales - muy mala	-0.0185***	0.00261
Provincia - 1102	0.0204***	0.002949
Provincia - 1103	0.0033	0.004568
Provincia - 1104	0.0935***	0.008668
Provincia - 1105	0.0164***	0.003419
Parámetro: σ		
Intercepto	-35.89***	0.000862
Gasto pagado	-0.0003***	0.000002
Alquiler	-0.0002***	0.000002
Estrato social - A, B o C	0.026***	0.000899
Global Deviance: 700 070.4		
AIC: 700 162.7		

Significancia estadística: *** 0.001 , ** 0.01 , * 0.05 , . 0.1

Tabla 7: Coeficientes del modelo GAMLSS para la estimación de μ_{ij} y σ_{ij} de la distribución $Binomial\ Negativa$ para la variable dependiente de número de habitaciones de las viviendas.
Elaboración propia a partir de estimaciones realizadas en R. Se han utilizado los factores de expansión de la ENAHO como pesos, con el propósito de considerar el diseño muestral dentro de la estimación del modelo.

El área correspondiente al ubigeo 110101 con el mayor tamaño de muestra obtuvo un promedio de 3.780 habitaciones, con un intervalo de confianza de 3.4704 y 4.0888. Por otro lado, para el área con ubigeo 110209 con una muestra de solo 5 viviendas se estimó un promedio de 2.097 habitaciones, con un intervalo de confianza más ancho de 1.6834 y 2.5107.

En este caso, según lo indicado en la Figura 19, se evidencia nuevamente un número elevado de áreas cuyo valor poblacional H_j no se encuentra dentro del intervalo de confianza al 95 % de $\hat{\hat{H}}_j$. Asimismo, al igual que el indicador de acceso a saneamiento mejorado,

los elevados tamaños de muestra de las áreas no aseguran obtener estimaciones cercanas al valor poblacional. Ello es el caso de las áreas a nivel distrital cuyos tamaños de muestra superan o aproximan las 100 viviendas, así como para el caso de las estimaciones de las provincias y todo el departamento.

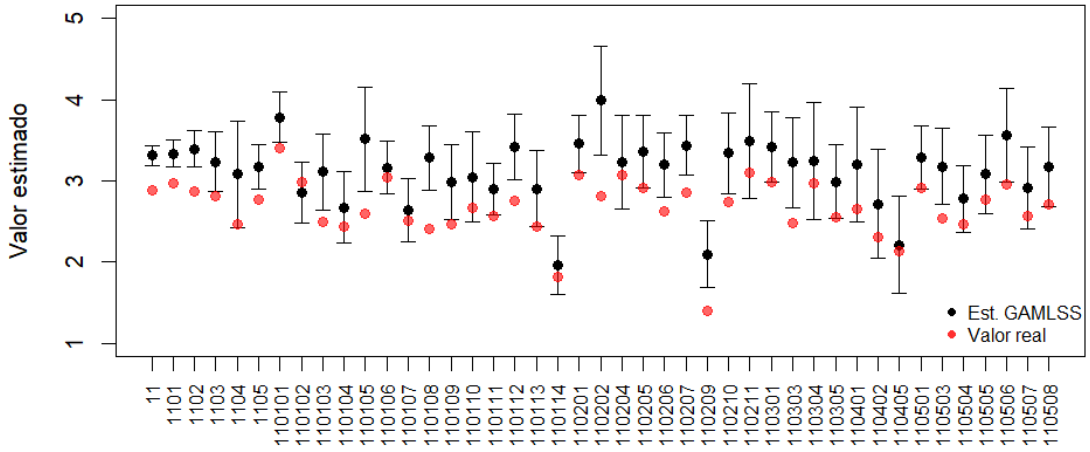


Figura 19: Intervalos de confianza al 95 % y contraste con valores poblacionales H_j del indicador de cantidad promedio de habitaciones de las viviendas según ubigeos.
Elaboración propia a partir de estimaciones realizadas en R.

5.2. Evaluación de la estimación de los indicadores

La sección anterior mostró las estimaciones GAMLSS de la proporción de viviendas que tienen servicios higiénicos conectados a una red pública de desagüe dentro de la vivienda y del promedio de habitaciones de las viviendas para los 37 distritos de Ica coberturados en la Encuesta Nacional de Hogares del año 2017. De igual manera, se realizaron las estimaciones de los indicadores para las cinco provincias y para todo el departamento. A continuación, se realizará la evaluación de dichas estimaciones, recurriendo a las estadísticas utilizadas en las simulaciones estadísticas del capítulo 4 y contrastándolas con las obtenidas mediante estimaciones directas[17], para lo cual se utilizarán los estimadores de Horvitz-Thompson bajo diseños de estratificación con conglomerados.

Las estadísticas agregadas, denotadas en la Tabla 9, indican que las estimaciones por GAMLSS presentan brechas relativas más reducidas entre $\hat{\hat{H}}_j$ y H_j. Estas diferencias son

[17]No se realiza la comparación con los estimadores de áreas pequeñas del modelo de Battese, Harter y Fuller (1988) ya que no se cuenta con información de las medias poblacionales de las covariables utilizadas.

N°	Ubigeo	n_j	$\hat{\bar{H}}_j$	IC (95 %)	N°	Ubigeo	n_j	$\hat{\bar{H}}_j$	IC (95 %)
Departamento									
1	11	1398	3.310	[3.1934 - 3.4265]					
Provincias									
2	1101	708	3.336	[3.1666 - 3.5058]	5	1104	36	3.082	[2.4242 - 3.7401]
3	1102	308	3.395	[3.1694 - 3.621]	6	1105	221	3.173	[2.8957 - 3.4506]
4	1103	125	3.237	[2.8694 - 3.6052]					
Distritos									
7	110101	285	3.780	[3.4704 - 4.0888]	26	110207	71	3.435	[3.0649 - 3.8044]
8	110102	68	2.861	[2.4866 - 3.2353]	27	110209	5	2.097	[1.6834 - 2.5107]
9	110103	21	3.110	[2.6439 - 3.5758]	28	110210	28	3.342	[2.847 - 3.8373]
10	110104	14	2.671	[2.2317 - 3.111]	29	110211	6	3.494	[2.79 - 4.1982]
11	110105	5	3.511	[2.876 - 4.1461]	30	110301	49	3.417	[2.9785 - 3.8555]
12	110106	100	3.164	[2.8415 - 3.4858]	31	110303	15	3.224	[2.6683 - 3.7802]
13	110107	22	2.643	[2.2588 - 3.027]	32	110304	24	3.245	[2.5223 - 3.9676]
14	110108	44	3.283	[2.8896 - 3.6765]	33	110305	37	2.991	[2.5369 - 3.4449]
15	110109	19	2.987	[2.5239 - 3.4495]	34	110401	18	3.202	[2.4935 - 3.9108]
16	110110	20	3.048	[2.4958 - 3.6011]	35	110402	7	2.713	[2.0436 - 3.3831]
17	110111	53	2.896	[2.5806 - 3.2122]	36	110405	11	2.209	[1.6114 - 2.8075]
18	110112	37	3.418	[3.0164 - 3.8199]	37	110501	75	3.289	[2.8991 - 3.6794]
19	110113	6	2.902	[2.436 - 3.3681]	38	110503	19	3.177	[2.7104 - 3.6432]
20	110114	14	1.970	[1.6098 - 2.3306]	39	110504	46	2.778	[2.3685 - 3.1875]
21	110201	91	3.456	[3.1031 - 3.8094]	40	110505	16	3.083	[2.6009 - 3.5651]
22	110202	6	3.987	[3.3223 - 4.6508]	41	110506	13	3.558	[2.9828 - 4.1335]
23	110204	19	3.235	[2.6615 - 3.8094]	42	110507	28	2.913	[2.4078 - 3.4185]
24	110205	35	3.359	[2.9093 - 3.8084]	43	110508	24	3.173	[2.6791 - 3.6664]
25	110206	47	3.198	[2.7993 - 3.5961]					

Tabla 8: Estimaciones a nivel distrital e intervalos de confianza para el indicador de cantidad promedio de habitaciones de la viviendas.
Elaboración propia a partir de estimaciones realizadas en R.

más evidentes a nivel distrital y provincial, y particularmente para la estimación del indicador de acceso a saneamiento.

Respecto a la comparación del error cuadrático medio, la estimación por GAMLSS genera en promedio indicadores con menor varianza para el indicador de acceso a saneamiento, empero para el indicador de cantidad de habitaciones la varianza es similar o mayor que la obtenida con los estimadores directos. Esto se debe a la mayor varianza estimada de los efectos aleatorios de las áreas, indicado anteriormente. Finalmente, las estadísticas agregadas también indican que las estimaciones GAMLSS generan indicadores cuyos intervalos de confianza al 95 % sí contienen el valor promedio poblacional, al menos en mayor medida que las estimaciones directas en el nivel distrital.

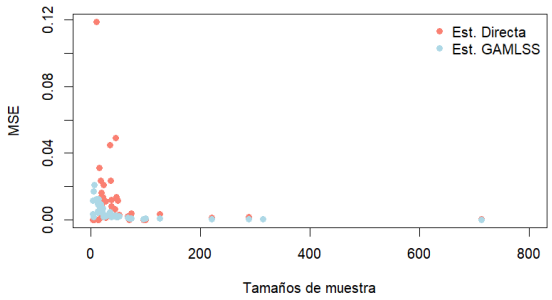
	Acceso a saneamiento		Cantidad de habitaciones	
	Est. Directa	Est. GAMLSS	Est. Directa	Est. GAMLSS
Departamento				
$\overline{MSE}$	0.0002	0.0001	0.0019	0.0035
$\overline{Diff}$	0.1867	0.1319	0.1419	0.1471
$\overline{IC}$	1.0000	1.0000	1.0000	1.0000
Provincias				
$\overline{MSE}$	0.0059	0.0011	0.0177	0.0377
$\overline{Diff}$	0.1920	0.1395	0.1704	0.1705
$\overline{IC}$	0.8000	1.0000	1.0000	0.8000
Distritos				
$\overline{MSE}$	0.0116	0.0050	0.0491	0.0635
$\overline{Diff}$	0.2714	0.1974	0.2190	0.1768
$\overline{IC}$	0.6216	0.3514	0.6757	0.4324

Tabla 9: Comparación de estadísticas agregadas de las estimaciones GAMLSS y estimaciones directas para la proporción de viviendas que cuentan con servicios higiénicos conectados a una red pública de desagüe y para la cantidad promedio de habitaciones de las viviendas en 37 distritos de Ica durante el año 2017. Elaboración propia a partir de estimaciones realizadas en R.

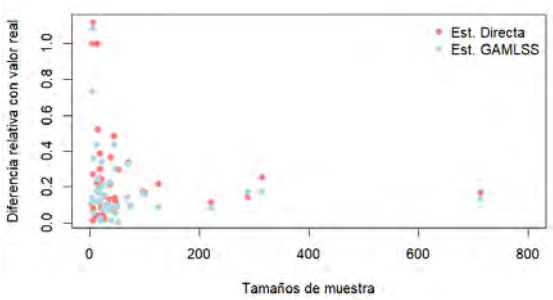
Al realizar este análisis según tamaños de muestra de las áreas, se evidencia que el menor error cuadrático medio de las estimaciones $\hat{\bar{H}}_j$ por GAMLSS es más evidente en las áreas con menor cantidad de observaciones particularmente para el indicador de acceso a saneamiento (ver Figuras 20 y 21). En este caso, se observa que luego de un tamaño de muestra mayor a 100, los errores cuadrático medio de las estimaciones por GAMLSS se asemejan a las estimaciones directas según diseño muestral.

Las diferencias relativas entre $\hat{\bar{H}}_j$ y H_j según tamaño de muestra no indican un patrón concreto. Sin embargo, en el caso del indicador de cantidad de habitaciones promedio de las viviendas, las diferencias relativas obtenidas mediante GAMLSS son menores a las obtenidas considerando solo el diseño muestral, especialmente en muestras pequeñas. Este resultado también es consistente con las simulaciones estadísticas realizadas para la variable numérica discreta, en el capítulo 4.

A pesar de lo anterior, según lo indicado en la Tabla 9, si bien la proporción de áreas cuyo valor promedio poblacional no se encuentra en el intervalo de confianza de $\hat{\bar{H}}_j$ de los GAMLSS es menor a la de las estimaciones directas de Horvitz-Thompson, su cifra es elevada. A nivel distrital, se ubica entre 35 % y 43 % dependiendo del indicador evaluado, y a nivel provincial y departamental la proporción de ubica entre 80 % y 100 % de las áreas.

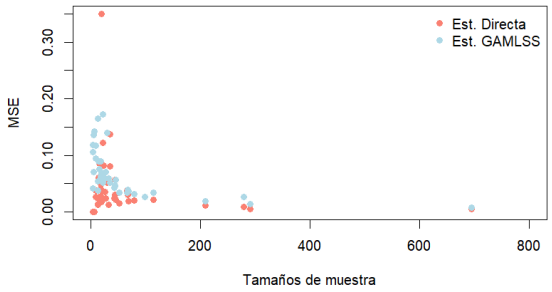


(a) $\widehat{MSE}(\hat{\bar{H}}_j)$

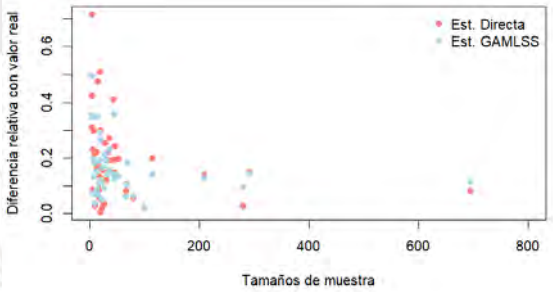


(b) $|\hat{\bar{H}}_j - H_j|/H_j$

Figura 20: Estadísticas comparativas desagregadas según tamaños de muestra y métodos de estimación para la proporción de viviendas que cuentan con servicios higiénicos conectados a una red pública de desagüe.
Elaboración propia a partir de estimaciones realizadas en R. Se ha omitido del gráfico la estimación departamental para evitar una desconfiguración de la escala del eje de las abscisas.



(a) $\widehat{MSE}(\hat{\bar{H}}_j)$



(b) $|\hat{\bar{H}}_j - H_j|/H_j$

Figura 21: Estadísticas comparativas desagregadas según tamaños de muestra y métodos de estimación para el promedio de habitaciones de las viviendas.
Elaboración propia a partir de estimaciones realizadas en R. Se ha omitido del gráfico la estimación departamental para evitar una desconfiguración de la escala del eje de las abscisas.

Esta diferencia respecto a lo obtenido en las simulaciones estadísticas del capítulo 4 se debe en parte al diseño muestral de la ENAHO. Como se ha indicado anteriormente, se trata de un esquema de estratificación y por conglomerados multietápico cuya población objetivo consiste de las viviendas particulares y sus residentes del área urbana y rural del Perú. En ese sentido, según la ficha técnica de la Encuesta Nacional de Hogares 2017, la estratificación se realiza considerando diferentes estratos geográficos en función de la cantidad de habitantes, asegurando una cobertura sufiente para los ámbitos urbano y rural. Por su lado, el muestreo multietápico considera el centro poblado, el conglomerado (área conformada por una o más manzanas) y la vivienda como unidades primarias, secundarias y terciarias de muestreo.

Según lo anterior, resulta evidente que el diseño muestral no toma en consideración

las demarcaciones territoriales (distritos, provincias o departamentos), lo cual genera que las estimaciones de los indicadores a nivel distrital se alejen de su valor poblacional, aun cuando los tamaños de muestra sean elevados. Esto motiva un análisis más a fondo de los GAMLSS bajo un contexto de áreas pequeñas considerando diferentes esquemas de muestreo y niveles de representatividad.

No obstante, una segunda posibilidad que explica las diferencias encontradas es que la Encuesta Nacional de Hogares no logra representar adecuadamente las estructuras urbano-rurales a nivel local del censo nacional. Ello se alinea a lo indicado en su ficha técnica, la cual menciona que el diseño muestral de la ENAHO brinda la suficiente cobertura para obtener estimaciones hasta para los 24 departamentos del Perú como dominios de estudio.

Al respecto, si se compara la estructura de la población de las cinco provincias de Ica según ámbito geográfico usando los datos del Censo Nacional 2017 y la ENAHO del mismo año, se encuentran diferencias relevantes. Por ejemplo, según el Censo Nacional en la provincia de Palpa (ubigeo 1104) el 57.1 % de su población residen en ámbitos rurales mientras que la ENAHO indica que dicha cifra es de 46.8 %, generándose una brecha de 10.2 puntos porcentuales. Las diferencias son evidentes aunque más reducidas en las provincias de Pisco (ubigeo 1105) y Chincha (ubigeo 1102) con brechas de 4.3 y 3.8 puntos porcentuales respectivamente. De esta manera, la ENAHO estaría subestimando el tamaño de la población rural a nivel provincial y, por ende, distrital. Ello podría explicar la distancia entre las estimaciones GAMLSS y los valores poblacionales obtenidos del censo, la cual se puede observar en las Figuras 16 y 19.

De hecho, este tipo de inconsistencias se pueden observar hasta en el nivel departamental, a pesar de lo indicado anteriormente respecto al diseño muestral de la ENAHO. Al realizar un contraste entre los valores censales y las estimaciones directas de Horvitz-Thompson de la cantidad promedio de habitaciones de las viviendas para los 24 departamentos del Perú y la Provincia Constitucional del Callao, se esperaría que los valores sean similares, ya que tanto la encuesta como censo se realizaron en el mismo periodo. No obstante, al observar la Figura 22, se aprecia un claro sesgo positivo a favor de las estadísticas de la ENAHO, particularmente para las viviendas del ámbito rural.

Este análisis indica que, para esta aplicación en particular, sería dificil obtener estimaciones que se aproximen a los valores del Censo Nacional usando la Encuesta Nacional de Hogares, tanto a nivel distrital, provincial o departamental, en parte por un aparente sesgo que provoca una brecha persistente entre valores poblacionales y muestrales particularmente en las áreas rurales.

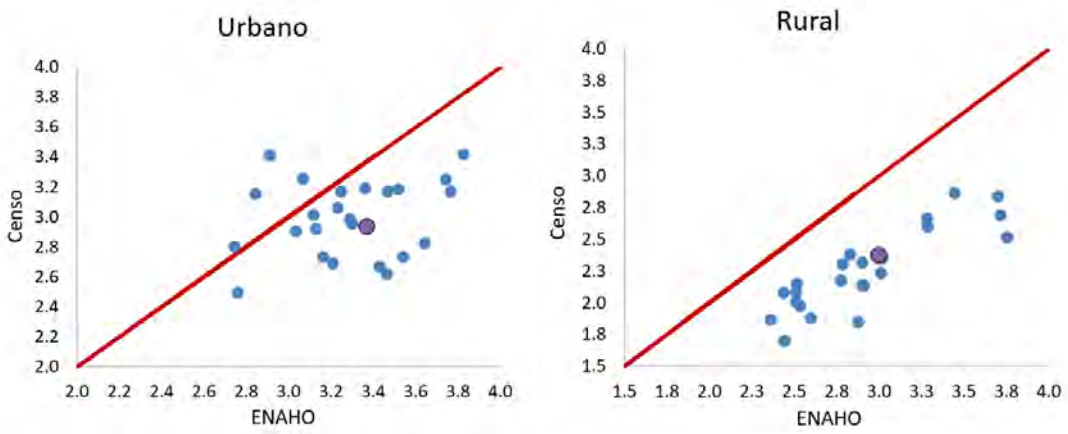


Figura 22: Contraste de indicadores de número promedio de habitaciones con valores poblaciones del Censo Nacional para los 24 departamentos del Perú y Callao, según ámbito geográfico.
Elaboración propia a partir de estimaciones realizadas en R. La ícono morado representa la estimación para el departamento de Ica

Sin embargo y a pesar de lo descrito líneas arriba, no se pueden ignorar las bondades de los GAMLSS demostradas anteriormento como el menor error cuadrático medio y la mayor exactitud de los indicadores cuando se compara con estimadores directos como los de Horvitz-Thompson. Respecto a la mayor precisión obtenida, en los siguientes gráficos de dispersión se muestra como el GAMLSS permite obtener indicadores más similares a los valores poblacionales del Censo Nacional para los distritos y provincias de Ica, al ubicarse más cerca a la línea de 45° que indica igualdad entre valor estimado y poblacional.

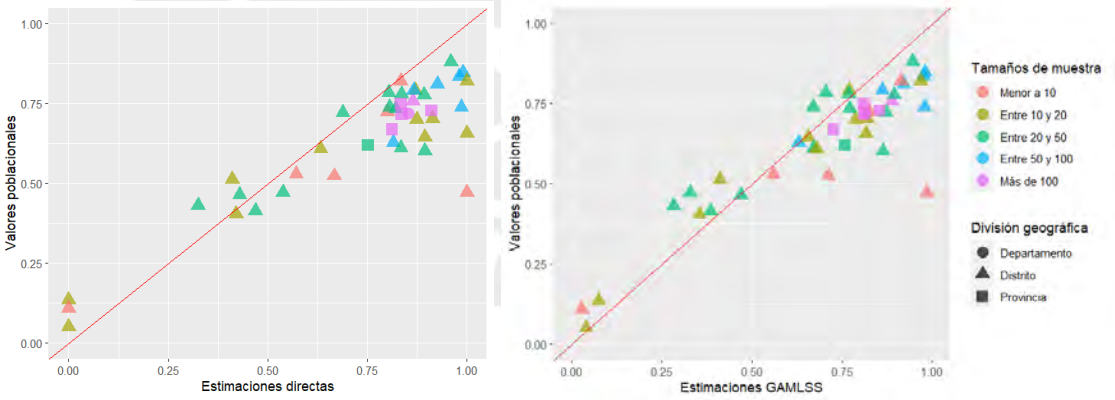


Figura 23: Contraste de indicadores de acceso a saneamiento con valores poblacionales del Censo Nacional para los distritos, las provincias y el departamento de Ica, según método de estimación y tamaño de muestra.
Elaboración propia a partir de estimaciones realizadas en R.

En el caso del indicador de acceso a saneamiento, el GAMLSS mejora la similitud de las estimaciones a los valores poblacionales especialmente para los distritos con menor tamaño de muestra y cuya estimación directa sobreestimaba su proporción de viviendas con acceso a redes públicas de alcantarillado dentro del hogar. Por el lado del indicador

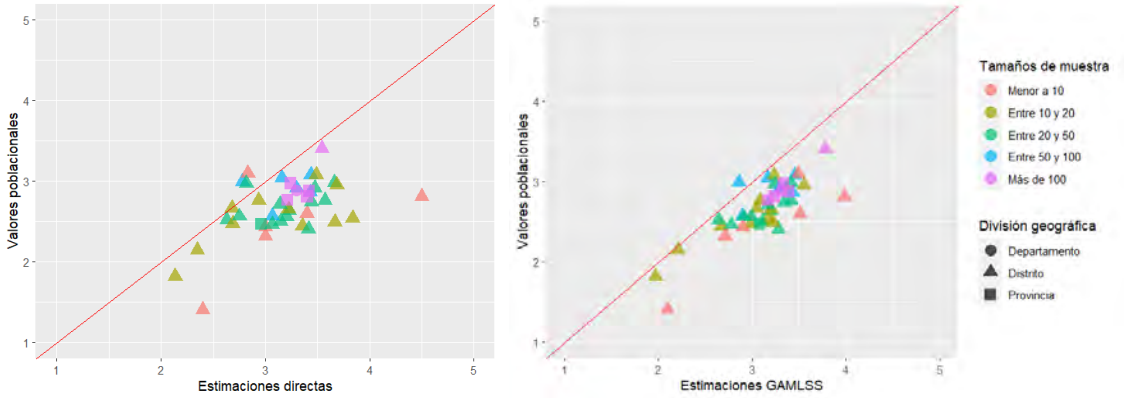


Figura 24: Contraste de indicadores de número promedio de habitaciones con valores poblacionales del Censo Nacional para los distritos, las provincias y el departamento de Ica, según método de estimación y tamaño de muestra.
Elaboración propia a partir de estimaciones realizadas en R.

de habitaciones, si bien el GAMLSS no elimina totalmente la brecha descrita anteriormente entre el Censo Nacional y la ENAHO, sí ayuda a reducirla, ya que bajo este método de estimación la nube de puntos se mueve hacia la izquierda del gráfico de dispersión, más cerca a la igualdad con los valores poblacionales.

Conclusiones

6.1. Conclusiones

La presente investigación ha buscado aportar al estado del arte de la estimación en áreas pequeñas, evaluando la robustez del uso de modelos aditivos de ubicación, escala y forma (GAMLSS) en este contexto y aplicándolo a datos reales de encuestas nacionales de Perú.

En línea con lo hallado por Mori y Ferrante (2023), las simulaciones estadísticas realizadas demuestran que las estimaciones en áreas pequeñas usando este tipo de modelos tienen errores cuadrático medio más pequeños cuando se comparan con estimaciones directas según diseño muestral, particularmente cuando las muestras de las áreas son de menor tamaño. También se ha encontrado que las estimaciones por GAMLSS tienden a ser más exactas respecto a los valores poblacionales, dependiendo de que tan heterogeneas sean las áreas: si la varianza de los efectos aleatorios de las áreas es elevada entonces no se encuentran diferencias importantes en la exactitud de las estimaciones mediante GAMLSS con las estimaciones directas.

Las simulaciones también han demostrado las bondades de los modelos aditivos de ubicación, escala y forma bajo diferentes formas de la distribución de la variable dependiente. Bajo el supuesto de una variable dependiente dicotómica con datos desbalanceados, las estimaciones por GAMLSS son más precisas y exactas si se comparan con estimaciones directas. Bajo el supuesto de una variable dependiente numérica discreta o continua no normal, el error cuadrático medio de las estimaciones de las áreas es más bajo pero pierden exactitud si las áreas son muy heterogeneas.

Asimismo, se ha contrastado el rendimiento de las estimaciones por GAMLSS con las obtenidas mediante el modelo de regresión lineal de error anidado de Battese, Harter y Fuller (1988). Las simulaciones demuestran que ambos métodos generan resultados generalmente similares, tanto en error cuadrático medio como en exactitud respecto a los valores poblacionales, siendo la única excepción el caso de variables continuos no normales con elevada heterogeneidad donde el método de BHF generó indicadores con mayor varianza. Ello se alinea a lo encontrado por Rojas-Perilla, Pannier, Schmid y Tzavidis (2020) quienes indican que bajo ausencia de normalidad de los efectos aleatorios de las áreas y los errores, las predicciones del modelo de Battese, Harter y Fuller (1988) son imprecisas. Por otro lado, es notable que bajo las simulaciones estadísticas el GAMLSS logra replicar la calidad de las predicciones del modelo de regresión lineal de error anidado, sin requerir las medias poblacionales de las covariables del modelo.

Finalmente, se han aplicado los GAMLSS a microdatos de la Encuesta Nacional de Hogares del año 2017, correspondientes a 37 distritos de la región de Ica en Perú. Con ello, se ha estimado puntualmente la proporción de viviendas que tienen servicios higiénicos conectados a una red pública de desagüe dentro de la vivienda y el promedio de habitaciones de las viviendas para los 37 distritos, considerando además estimaciones para sus cinco provincias y para todo el departamento. Asimismo, se han obtenido sus intervalos de confianza al 95 % a partir de las estimaciones de sus errores cuadrático medio.

Siguiendo con lo obtenido en las simulaciones estadísticas, estas estimaciones demostraron tener menores brechas entre los indicadores estimados y los obtenidos por el Censo Nacional del 2017 que las estimaciones directas a partir del diseño muestral de la encuesta. Asimismo, para el indicador de acceso a servicios higiénicos, se obtuvieron errores cuadráticos medio más reducidos. Caso contrario se observó para el indicador de habitaciones, aunque ello estaría explicado por una limitada bondad de ajuste del modelo planteado.

No obstante, a pesar que los resultados de la aplicación por GAMLSS fueron relativamente mejores que las estimaciones directas, se ha encontrado un numero importante de áreas donde el intervalo de confianza al 95 % de los indicadores estimados no contienen el valor poblacional del censo nacional. Ello podría deberse a la elevada heterogeneidad de las áreas o a las covariables elegidas para modelizar las variables dependientes, pero también a un posible sesgo en los datos de la encuesta nacional que no permite replicar la estructura urbano-rural recogida en el censo. Lo anterior también podría deberse a que la encuesta nacional no considera las demarcaciones territoriales (distritos, provincias o departamentos) en su diseño muestral.

6.2. Sugerencias para investigaciones futuras

Con el propósito de mejorar la aplicación de los GAMLSS a la estimación en áreas pequeñas, se sugiere estudiar la relajación del supuesto de ausencia de autocorrelación espacial de los efectos aleatorios de la áreas con el propósito de aumentar la robustes del error cuadrático medio de los indicadores. Asimismo, se sugiere estudiar el impacto que tiene una especificación errónea de la variable dependiente (sea en su distribución o variables explicativas) en la robustez de los GAMLSS, ya que se ha encontrado que una limitada bondad de ajuste tiene incidencias en el error cuadrático medio de sus estimaciones.

Por otro lado, se recomienda analizar en mayor detalle las bondades de los GAMLSS bajo diferentes tamaños de muestra y diseños muestrales, ya que se ha encontrado que la baja representatividad de las áreas pequeñas (como distritos) en la muestra puede ser un obstáculo para la presición de este método. Finalmente, se sugiere estudiar en mayor

detalle la brecha identificada entre los indicadores de infraestructura de las viviendas del Censo Nacional y de la Encuesta Nacional de Hogares del 2017, considerando que esta última debería proveer información suficiente para obtener estadísticas similares a los valores poblacionales.

Referencias

Battese, G., Harter, R., & Fuller, W. (1988). Components Model for Prediction of CountyCrop Areas Using Survey and Satellite Data. *Journal of the American Statistical Association*, 83, 28-36.

Dunn, P. K., & Smyth, G. K. (1996). Randomized Quantile Residuals. *Journal of Computational and Graphical Statistics*, 5(3), 236?244.

Fay, R., & Herriot, R. (1979). Estimation of income from small places: An application of James-Steinprocedures to census data. *Journal of the American Statistical Association*, 74, 269-277.

Efron, B. (1979). Bootstrap Methods: Another Look at the Jackknife. *The Annals of Statistics*, 7 (1) 1 - 26.

Efron, B., & Tibshirani, R. J. (1993). An introduction to the bootstrap. New York: Chapman & Hall

Elbers, C., Lanjouw.J.O., & Lanjouw, P. (2003). Micro-level estimation of poverty and inequality. *Econometrika*, 71, 355-364.

Gonzalez-Manteiga, W., Lombardía, M.J., Morales, D. & Santamaría, L. (2008). Bootstrap mean squared error of a small-area EBLUP. *Journal of Statistical Computation and Simulation*, Vol. 78, 5, 443-462.

Graf, M., Marín, J. M., & Molina, I. (2019). A generalized mixed model for skewed distributions applied to small area estimation. *Test*, 28, 565-597.

Graubar, Barry I. & Korn, Edward L. (2002). Inference for Superpopulation Parameters Using Sample Surveys. *Statistical Science*, 17(1), 73-96.

Guadarrama, M., Molina, I., & Rao, J. (2014). A comparison of small area estimation methods for poverty mapping. *STATISTICS IN TRANSITION new series and SURVEY METHODOLOGY. Joint Issue: Small Area Estimation*, 17(1), 41-66.

Haro, M. E. (2022). Un modelo Fay-Herriot espacial para la predicción del porcentaje de niños con anemia y riesgo de retraso del crecimiento en distritos no encuestados y en distritos con pocas observaciones disponibles. Lima: Pontificia Universidad Católica del Perú. Obtenido de https://tesis.pucp.edu.pe/repositorio/handle/20.500.12404/24341.

Lee, K., Joo, S., Baik, H., Han, S. & In, J. (2020). Unbalanced data, type II error, and nonlinearity in predicting M&A failure. *Journal of Business Research*,Volume 109, 271-287.

López, J., & Francés, F. (2013). Non-stationary flood frequency analysis in continental Spanish rivers, using climate and reservoir indices as external covariates. *Hydrology and Earth System Sciences*, 17(8), 3189-3203.

Molina, I., & Rao, J. (2010). Small area estimation of poverty indicators. *The Canadian Journal of Statistics*, 38(3), 369-385.

Mori, L., & Ferrante, M. R. (2023). Small Area Estimation Under Unit-Level Generalized Additive Models for Location, Scale and Shape. *Psychometrika*, 74(2), 191-210.

Rao, J., & Hussain Chounhry, C.(1995). Small Area Estimation: Overview and Empirical Study. In B. G. Cox, D. A. Binder, B. N. Chinnappa, A. Christianson, M. J. Colledge, and P. S. Kott, Business Surver Methods (pp. 527-542). *John Wiley & Sons, Inc*.

Rigby, R. A., & Stasinopoulos, D. M. (1996). A semi-parametric additive model for variance heterogeneity. *Statistics and Computing*, 6, 57-65.

Rigby, R. A., Stasinopoulos, D. M., Heller, G. Z., & De Bastiani, F. (2019) Distributions for modeling location, scale, and shape: Using GAMLSS in R, Chapman and Hall.

Rojas-Perilla, N., Pannier, S., Schmid, T., & Tzavidis, N.(2020). Data-driven transformations in small area estimation. *J. R. Statist. Soc. A*, 121-148.

Stasinopoulos, D. M., & Rigby, R. A. (2007). Generalized Additive Models for Location Scale and Shape (GAMLSS) in R. *Journal of Statistical Software*, 23(7), 1-46.

Stasinopoulos, M. D., Rigby, R. A., & Bastiani, F. D. (2018). GAMLSS: A distributional regression approach. *Statistical Modelling*, 18(3-4), 248-273.

Stevens, J. P. (1984). Outliers and influential data points in regression analysis. *Psychological Bulletin*, 95(2), 334?344.

Voudouris, V., Gilchrist, R., Rigby, R., Sedgwick, J., & Stasinopoulos, D. (2012). Modelling skewness and kurtosis with the BCPE density in GAMLSS. *Journal of Applied Statistics*, 39(6), 1279-1293.

Apéndices

En la presente sección se muestran los códigos de R utilizados en la aplicación de GAMLSS para estimación de los indicadores de infraestructura de hogares a nivel distrital.

7.1. Códigos de R para la construcción de la base de datos

```
library(haven)
library(ggplot2)
library(gridExtra)

####IMPORTACION DE BASE DE DATOS####

enaho01_2017_100=read_dta("enaho01-2017-100.dta")

#Generacion de variable de departamento a partir de ubigeo
enaho01_2017_100$departamento=substr(enaho01_2017_100$ubigeo,1,2)

####FUSION CON OTROS MODULOS####

#Modulo: Sumaria
sumaria_2017=read_dta("sumaria-2017.dta")
sumaria_2017=sumaria_2017[,c("conglome","vivienda","hogar","pobreza",
"mieperho","estrsocial")]
enaho01_2017_100=merge(enaho01_2017_100,sumaria_2017,all.x = TRUE,
by=c("conglome","vivienda","hogar"))

#Modulo: Educacion
enaho01a_2017_300=read_dta("enaho01a-2017-300.dta")
enaho01a_2017_300=
enaho01a_2017_300[,c("conglome","vivienda","hogar","codperso","p301a")]
enaho01a_2017_300=
aggregate(p301a~conglome+vivienda+hogar,data=enaho01a_2017_300,FUN="max")
enaho01_2017_100=merge(enaho01_2017_100,enaho01a_2017_300,
all.x = TRUE,by=c("conglome","vivienda","hogar"))

#Modulo: Percepcion del gobierno
enaho01b_2017_1=read_dta("enaho01b-2017-1.dta")
enaho01b_2017_1=
enaho01b_2017_1[,c("conglome","vivienda","hogar","codperso","p2a1_3","p2a1_4")]
enaho01_2017_100=merge(enaho01_2017_100,enaho01b_2017_1,
all.x = TRUE,by=c("conglome","vivienda","hogar"))

####GENERACION DE VARIABLES####
```

```
##Variables dependientes

#acceso a saneamiento por red pública dentro del hogar
enaho01_2017_100$saneamiento=enaho01_2017_100$p111a==1

#número de habitaciones
enaho01_2017_100$habitaciones=enaho01_2017_100$p104

##Covariables

#Ubicacion geografica - sector urbano
enaho01_2017_100$urbano=enaho01_2017_100$estrato>=1 & enaho01_2017_100$estrato<=5

#Monto de alquiler
enaho01_2017_100$alquiler=
rowSums(cbind(enaho01_2017_100$p105b,enaho01_2017_100$p106),na.rm=TRUE)

#Gasto pagado total mensual pagado
enaho01_2017_100$gasto_pagado=enaho01_2017_100$p117t2

#Pared de cemento
enaho01_2017_100$pared_cemento=enaho01_2017_100$p102==1

#Piso de tierra
enaho01_2017_100$piso_tierra=enaho01_2017_100$p103==6

#Gasto para servicios de agua
enaho01_2017_100$gasto_agua=enaho01_2017_100$p1172_01

#Vivienda propia
enaho01_2017_100$vivienda_propia=enaho01_2017_100$p105a==2

#Estrato social a,b o c
enaho01_2017_100$estr_social_alto=enaho01_2017_100$estrsocial==1 |
enaho01_2017_100$estrsocial==2 | enaho01_2017_100$estrsocial==3
enaho01_2017_100$estrsocial=factor(enaho01_2017_100$estrsocial)

#Número de miembros del hogar
enaho01_2017_100$m_hogar=enaho01_2017_100$mieperho

#Maximo nivel educativo alcanzado por algún miembro del hogar
enaho01_2017_100$niv_edu=enaho01_2017_100$p301a
enaho01_2017_100$niv_edu=ifelse(enaho01_2017_100$niv_edu==1,1,
ifelse(enaho01_2017_100$niv_edu==2 | enaho01_2017_100$niv_edu==3,2,
ifelse(enaho01_2017_100$niv_edu==4 | enaho01_2017_100$niv_edu==5,3,
```

```
ifelse(enaho01_2017_100$niv_edu==6 | enaho01_2017_100$niv_edu==7 |
enaho01_2017_100$niv_edu==9 ,4,
ifelse(enaho01_2017_100$niv_edu==8 | enaho01_2017_100$niv_edu==10,5,
ifelse(enaho01_2017_100$niv_edu==11,6,7))))))

enaho01_2017_100$niv_edu=factor(enaho01_2017_100$niv_edu,levels=c(1:7),
labels = c("Sin nivel","Educacion inicial",
"Primaria","Secundaria","Superior",
"Postgrado","Especial"))

enaho01_2017_100$niv_edu_sup=enaho01_2017_100$niv_edu=="Superior" |
enaho01_2017_100$niv_edu=="Postgrado"

#Percepción del jefe del hogar respecto a los gobiernos locales
enaho01_2017_100$percepcion=round(rowMeans(enaho01_2017_100[,c("p2a1_3","p2a1_4")]),0)
enaho01_2017_100$percepcion=factor(enaho01_2017_100$percepcion,
levels=c(1:5),
labels = c("Muy buena","Buena","Mala","Muy mala","No sabe"))

enaho01_2017_100$percepcion_muy_mala=enaho01_2017_100$percepcion=="Muy mala"

####ANALISIS PREVIO####

#Mantener solo datos de Ica
enaho01_2017_100_ICA=enaho01_2017_100[enaho01_2017_100$departamento=="11",]

#Mantener solo datos de casas independientes, departamentos,
#viviendas en quinta y viviendas en casa de vecindad
enaho01_2017_100_ICA=enaho01_2017_100_ICA[enaho01_2017_100_ICA$p101>=1 &
enaho01_2017_100_ICA$p101<=4,]

attach(enaho01_2017_100_ICA)

#Acceso a saneamiento
summary(saneamiento)
proportions(table(saneamiento))

#FIGURAS
datos_figuras_sanea=data.frame(saneamiento,urbano,piso_tierra,gasto_agua,
niv_edu_sup,estr_social_alto,percepcion_muy_mala)
datos_figuras_sanea=datos_figuras_sanea[complete.cases(datos_figuras_sanea),]
datos_figuras_sanea$saneamiento=factor(datos_figuras_sanea$saneamiento,
levels=c(TRUE,FALSE),labels = c("Sí","No"))
datos_figuras_sanea$urbano=factor(datos_figuras_sanea$urbano,
levels=c(TRUE,FALSE),labels = c("Urbano","Rural"))
datos_figuras_sanea$piso_tierra=factor(datos_figuras_sanea$piso_tierra,
```

```
levels=c(TRUE,FALSE),
labels = c("Piso de tierra","Otros pisos"))
datos_figuras_sanea$niv_edu_sup=factor(datos_figuras_sanea$niv_edu_sup,
levels=c(TRUE,FALSE),
labels = c("Edu. superior","No edu. superior"))
datos_figuras_sanea$estr_social_alto=factor(datos_figuras_sanea$estr_social_alto,
levels=c(TRUE,FALSE),
labels = c("Estr. social alto","Estr. social bajo"))
datos_figuras_sanea$percepcion_muy_mala=factor(datos_figuras_sanea$percepcion_muy_mala,
levels=c(TRUE,FALSE),
labels = c("Mala percepción","Buena percepción"))

#Funciones de apoyo para figuras
comp_pct <- function(count, group) {
count / tapply(count, group, sum)[group]
}

get_legend<-function(myggplot){
tmp <- ggplot_gtable(ggplot_build(myggplot))
leg <- which(sapply(tmp$grobs, function(x) x$name) == "guide-box")
legend <- tmp$grobs[[leg]]
return(legend)
}

plot_sanea1=ggplot(data=datos_figuras_sanea,
aes(x=urbano,fill=saneamiento)) + geom_bar(position = "fill") +
ylab("Porcentaje") + xlab(NULL) +
labs(fill="Saneamiento") +
geom_text(aes(label = after_stat(scales::percent(comp_pct(count, x))),
y = after_stat(count)), stat = "count", position = position_fill(vjust=0.5))

plot_sanea2=ggplot(data=datos_figuras_sanea,
aes(x=estr_social_alto,fill=saneamiento)) + geom_bar(position = "fill") +
ylab("Porcentaje") + xlab(NULL) + labs(fill="Saneamiento") +
geom_text(aes(label = after_stat(scales::percent(comp_pct(count, x))),y = after_stat(count))
, stat = "count", position = position_fill(vjust=0.5)) + theme(legend.position="none")

plot_sanea3=ggplot(data=datos_figuras_sanea,
aes(x=piso_tierra,fill=saneamiento)) + geom_bar(position = "fill") + ylab("Porcentaje") + xlab(NULL) +
labs(fill="Saneamiento") +
geom_text(aes(label = after_stat(scales::percent(comp_pct(count, x))),y = after_stat(count))
, stat = "count", position = position_fill(vjust=0.5)) + theme(legend.position="none")

plot_sanea4=ggplot(data=datos_figuras_sanea,
aes(x=niv_edu_sup,fill=saneamiento)) +
geom_bar(position = "fill") + ylab("Porcentaje") + xlab(NULL) + labs(fill="Saneamiento") +
```

```
geom_text(aes(label = after_stat(scales::percent(comp_pct(count, x))),y = after_stat(count))
, stat = "count", position = position_fill(vjust=0.5)) + theme(legend.position="none")

plot_sanea5=ggplot(data = datos_figuras_sanea, aes(x = saneamiento, y = gasto_agua)) +
geom_boxplot() + ylab("Gasto mensual por consumo de agua") + xlab("Saneamiento")

legend_sanea=get_legend(plot_sanea1)
plot_sanea1=plot_sanea1 + theme(legend.position="none")
grid.arrange(plot_sanea5,plot_sanea1, plot_sanea2, plot_sanea3, plot_sanea4,legend_sanea,
ncol=4, nrow = 2, layout_matrix = rbind(c(1,2,3,6), c(1,4,5,6)),widths=c(2.7,2.7,2.7,0.9))

#Numero de habitaciones
summary(habitaciones)
proportions(table(habitaciones))

datos_figuras_habi=data.frame(habitaciones,niv_edu_sup,alquiler,gasto_pagado,vivienda_propia,m_hogar,
estr_social_alto)
datos_figuras_habi=datos_figuras_habi[complete.cases(datos_figuras_habi),]
datos_figuras_habi$niv_edu_sup=factor(datos_figuras_habi$niv_edu_sup,
levels=c(TRUE,FALSE),
labels = c("Tiene edu. sup","No tiene edu. sup"))
datos_figuras_habi$vivienda_propia=factor(datos_figuras_habi$vivienda_propia,
levels=c(TRUE,FALSE),
labels = c("Viv. propia","No viv. propia"))
datos_figuras_habi$estr_social_alto=factor(datos_figuras_habi$estr_social_alto,
levels=c(TRUE,FALSE),
labels = c("Estr. social alto","Estr. social bajo"))
datos_figuras_habi$habitaciones[datos_figuras_habi$habitaciones>=5]=5
datos_figuras_habi$habitaciones=factor(datos_figuras_habi$habitaciones,
levels=c(1:5),labels = c("1","2","3","4","Más de 4"))
proportions(table(datos_figuras_habi$habitaciones))

plot_habi1=ggplot(data = datos_figuras_habi, aes(x = habitaciones, y = alquiler)) +
geom_boxplot() + ylab("Monto de alquiler mensual") + xlab("Habitaciones")

plot_habi2=ggplot(data = datos_figuras_habi, aes(x = habitaciones, y = gasto_pagado)) +
geom_boxplot() + ylab("Gasto mensual total") + xlab("Habitaciones")

plot_habi3=ggplot(data=datos_figuras_habi,
aes(x=estr_social_alto,fill=habitaciones)) + geom_bar(position = "fill") +
ylab("Porcentaje") + xlab(NULL) + labs(fill="Habitaciones") +
geom_text(aes(label = after_stat(scales::percent(comp_pct(count, x))),
y = after_stat(count)) ,stat = "count", position = position_fill(vjust=0.5))

plot_habi4=ggplot(data=datos_figuras_habi,
aes(x=vivienda_propia,fill=habitaciones)) + geom_bar(position = "fill") +
```

```
ylab("Porcentaje") + xlab(NULL) + labs(fill="Habitaciones") +
geom_text(aes(label = after_stat(scales::percent(comp_pct(count, x))),y = after_stat(count))
,stat = "count", position = position_fill(vjust=0.5)) + theme(legend.position="none")

plot_habi5=ggplot(data = datos_figuras_habi, aes(x = habitaciones, y = m_hogar)) +
geom_boxplot() + ylab("Número de miembros del hogar") + xlab("Habitaciones")

legend_habi=get_legend(plot_habi3)
plot_habi3=plot_habi3 + theme(legend.position="none")
grid.arrange(plot_habi5,plot_habi1,plot_habi2, plot_habi3, plot_habi4, legend_habi,
ncol=4, nrow = 2, layout_matrix = rbind(c(1,2,4,6), c(1,3,5,6)),widths=c(2.7,2.7,2.7,0.9))

detach(enaho01_2017_100_ICA)

####LIMPIEZA DE BASE DE DATOS PARA ANÁLISIS####

lista_variables=c("conglome", "vivienda", "hogar", "ubigeo", "estrato",
"saneamiento", "habitaciones","urbano", "pared_cemento", "piso_tierra",
"vivienda_propia", "gasto_agua","niv_edu_sup","percepcion_muy_mala","estr_social_alto",
"alquiler", "gasto_pagado","m_hogar", "estrsocial",
"longitud", "latitud", "factor07")

BBDDENAHO2017=enaho01_2017_100_ICA[,lista_variables]
BBDDENAHO2017=BBDDENAHO2017[order(BBDDENAHO2017$ubigeo),]

save(BBDDENAHO2017,file="BBDDENAHO2017.Rda")
```

7.2. Códigos de R de GAMLSS para la estimación de indicadores de acceso a saneamiento de las viviendas

```
library(haven)
library(readxl)
library(gamlss.dist)
library(gamlss.ggplots)
library(gamlss)
library(sae)
library(dplyr)
library(sampling)
library(survey)
library(ggplot2)
library(collapse)
library(gridExtra)

####CARGA DE DATOS#####
```

```
setwd("C:/Users/hans_/Desktop/maestria/PUCP/Tesis 2")

load("BBDDENAHO2017.Rda")
DatosCenso2017=read_excel("DatosCenso2017.xlsx",
col_types = c("text", "numeric", "numeric","numeric"),
sheet = "Ica")
DatosCenso2017_prov=read_excel("DatosCenso2017.xlsx",
col_types = c("text", "numeric", "numeric","numeric"),
sheet = "Ica_prov")

DatosCenso2017_depa=read_excel("DatosCenso2017.xlsx",
col_types = c("text", "numeric", "numeric","numeric"),
sheet = "Ica_depa")

prom_Y_area=DatosCenso2017$saneamiento
prom_Y_area_prov=DatosCenso2017_prov$saneamiento
prom_Y_area_depa=DatosCenso2017_depa$saneamiento

Nj=DatosCenso2017$Nj
Nj_prov=DatosCenso2017_prov$Nj
Nj_depa=DatosCenso2017_depa$Nj

ubigeo=rep(as.numeric(DatosCenso2017$ubigeo),times=Nj)
J=length(as.numeric(DatosCenso2017$ubigeo))

BBDDENAHO2017$saneamiento=as.logical(BBDDENAHO2017$saneamiento)
BBDDENAHO2017$provincia=substr(BBDDENAHO2017$ubigeo,1,4)

BBDD=BBDDENAHO2017[complete.cases(BBDDENAHO2017),]

#incorporacion de dummies de provincia
BBDD$provincia_1102=BBDD$provincia=="1102"
BBDD$provincia_1103=BBDD$provincia=="1103"
BBDD$provincia_1104=BBDD$provincia=="1104"
BBDD$provincia_1105=BBDD$provincia=="1105"

#GLM se usa para identificar valores extremos
modelo_glm=glm(saneamiento~urbano+pared_cemento+piso_tierra+vivienda_propia+gasto_agua
+niv_edu_sup+percepcion_muy_mala+estr_social_alto+alquiler
+provincia_1102+provincia_1103+provincia_1104+provincia_1105,
family = binomial(link="logit"),data = BBDD, weights=factor07)

summary(modelo_glm)
plot(cooks.distance(modelo_glm))

nj=as.numeric(table(BBDD$ubigeo))
```

```
nj_prov=as.numeric(table(BBDD$provincia))
nj_depa=length(BBDD$provincia)

tasa_rep_nj=ceiling(Nj/nj)

hist(prom_Y_area,xlim = c(0,1),main = NULL,
xlab="Distribución de frecuencias de la proporción de acceso a saneamiento",
ylab="Frencuencias") ; summary(prom_Y_area)
boxplot(nj,xlab="Tamaños de muestra de los distritos de Ica",horizontal = TRUE)
summary(nj)

#####ESTIMACION DIRECTA CON PROMEDIOS#######
BBDD$ubigeo=as.numeric(BBDD$ubigeo)

#distrital
diseño_muestral=svydesign(ids=~conglome,weights=~factor07,strata=~estrato,data = BBDD)
est_indicador=svyby(~saneamiento,~ubigeo,diseño_muestral,svymean)$saneamientoTRUE
var_indicador=(svyby(~saneamiento,~ubigeo,diseño_muestral,svymean)$se.saneamientoTRUE)^2
MSE_directo=var_indicador

#IC
IC_sup_directo=est_indicador+1.96*sqrt(var_indicador) ; IC_sup_directo[IC_sup_directo>1]=1
IC_inf_directo=est_indicador-1.96*sqrt(var_indicador) ; IC_inf_directo[IC_inf_directo<0]=0

#contraste de resultados con medias reales
Diff_directo=abs(est_indicador-prom_Y_area)/prom_Y_area ; mean(Diff_directo)
NoDentroIC_directo=prom_Y_area<IC_inf_directo | prom_Y_area>IC_sup_directo
mean(NoDentroIC_directo)

#provincial
est_indicador_prov=svyby(~saneamiento,~provincia,diseño_muestral,svymean)$saneamientoTRUE
var_indicador_prov=(svyby(~saneamiento,~provincia,diseño_muestral,svymean)$se.saneamientoTRUE)^2
MSE_directo_prov=var_indicador_prov

#IC
IC_sup_directo_prov=est_indicador_prov+1.96*sqrt(var_indicador_prov)
IC_sup_directo_prov[IC_sup_directo_prov>1]=1

IC_inf_directo_prov=est_indicador_prov-1.96*sqrt(var_indicador_prov)
IC_inf_directo_prov[IC_inf_directo_prov<0]=0

#contraste de resultados con medias reales
Diff_directo_prov=abs(est_indicador_prov-prom_Y_area_prov)/prom_Y_area_prov
mean(Diff_directo_prov)
NoDentroIC_directo_prov=prom_Y_area_prov<IC_inf_directo_prov | prom_Y_area_prov>IC_sup_directo_prov
mean(NoDentroIC_directo_prov)
```

```
#departamental
est_indicador_depa=coef(svymean(~saneamiento,diseño_muestral))[2]
var_indicador_depa=SE(svymean(~saneamiento,diseño_muestral))[2]^2
MSE_directo_depa=var_indicador_depa

#IC
IC_sup_directo_depa=est_indicador_depa+1.96*sqrt(var_indicador_depa)
IC_sup_directo_depa[IC_sup_directo_depa>1]=1

IC_inf_directo_depa=est_indicador_depa-1.96*sqrt(var_indicador_depa)
IC_inf_directo_depa[IC_inf_directo_depa<0]=0

#contraste de resultados con medias reales
Diff_directo_depa=abs(est_indicador_depa-prom_Y_area_depa)/prom_Y_area_depa
mean(Diff_directo_depa)
NoDentroIC_directo_depa=prom_Y_area_depa<IC_inf_directo_depa | prom_Y_area_depa>IC_sup_directo_depa
mean(NoDentroIC_directo_depa)

#####ESTIMACION GAMLSS#####
t_old=Sys.time()

set.seed(432)

#Modelización del parámetro de la distribución de la variable binaria
modelo_gamlss=gamlss(saneamiento~urbano+pared_cemento+piso_tierra+vivienda_propia+gasto_agua
+niv_edu_sup+percepcion_muy_mala+estr_social_alto+alquiler
+provincia_1102+provincia_1103+provincia_1104+provincia_1105
+re(random=~1|ubigeo),
family = BI(mu.link="logit"),data = BBDD, weights=factor07,
control = gamlss.control(n.cyc = 40,c.crit = 0.01))

summary(modelo_gamlss)

plot(modelo_gamlss)

grid.arrange(
resid_mu(modelo_gamlss,title="Residuals vs Fitted Values",value = 3,annotate = FALSE),
resid_index(modelo_gamlss,title="Residuals vs Index",value = 3,annotate = FALSE),
resid_density(modelo_gamlss,title="Density Estimate"),
resid_qqplot(modelo_gamlss, value = 3, points.col = "steelblue4",
line.col = "darkgray", check_overlap = TRUE, title="Q-Q Plot"))

#Test de normalidad de errores (randomized-quantile residuales)
shapiro.test(resid(modelo_gamlss))
ks.test(resid(modelo_gamlss),"pnorm")
```

```
pi_hat_area=aggregate(pi_hat,list(BBDD$ubigeo), FUN=mean)$x
names(pi_hat)=BBDD$ubigeo

#Estimación de indicador - según método de Graf, Marin y Molina (2019)
#Simulacion de datos sintéticos de Y y estimación de indicador
L=200

datos_faltantes=Nj-nj
tasa_rep_nj_ajust=datos_faltantes/nj
Hj_simMC=NULL
Hj_simMC_prov=NULL
Hj_simMC_depa=NULL

for (k in 1:L) {

datos_sim_Y=mapply(rbinom,prob=pi_hat,n=rep(tasa_rep_nj_ajust,times=nj),
SIMPLIFY=TRUE,size=1)

id_area_faltantes=as.numeric(rep(names(datos_sim_Y), lengths(datos_sim_Y)))
datos_sim_Y_faltantes=unlist(datos_sim_Y,use.names = FALSE)
datos_sim_faltantes=data.frame(datos_sim_Y=datos_sim_Y_faltantes,id_area=id_area_faltantes)
datos_sim_final=rbind(data.frame(datos_sim_Y=BBDD$saneamiento,
id_area=BBDD$ubigeo),datos_sim_faltantes)
datos_sim_final$prov=floor(datos_sim_final$id_area/100)

Hj_sim=fmean(datos_sim_final[,c(1,2)],datos_sim_final$id_area)$datos_sim_Y
Hj_simMC=cbind(Hj_simMC,Hj_sim)

Hj_sim_prov=fmean(datos_sim_final[,c(1,3)],datos_sim_final$prov)$datos_sim_Y
Hj_simMC_prov=cbind(Hj_simMC_prov,Hj_sim_prov)

Hj_sim_depa=mean(datos_sim_final$datos_sim_Y)
Hj_simMC_depa=cbind(Hj_simMC_depa,Hj_sim_depa)
}

est_gamlss=as.numeric(rowMeans(Hj_simMC))
est_gamlss_prov=as.numeric(rowMeans(Hj_simMC_prov))
est_gamlss_depa=as.numeric(rowMeans(Hj_simMC_depa))

#Estimacion de MSE en GAMLSS

set.seed(432)
B=200

beta_est=coef(modelo_gamlss)[1:14] # coeficientes beta
```

```
gamma_est=as.numeric(unlist(coef(getSmo(modelo_gamlss)))) # efectos aleatorios de las areas
var_gamma=as.numeric(VarCorr(getSmo(modelo_gamlss))[1])
m_var_cov=diag(var_gamma,J)

desv_cuadra_boot=NULL
dist_gamlss_boot=NULL

desv_cuadra_boot_prov=NULL
dist_gamlss_boot_prov=NULL

desv_cuadra_boot_depa=NULL
dist_gamlss_boot_depa=NULL

for (b in 1:B) {

#Simulación nueva de gamma
t_sim=rnorm(J)
gamma_sim=sqrt(m_var_cov)%*%t_sim
gamma_sim_x=rep(gamma_sim,times=nj)

#Simulación nueva de pi_hat (inversa de función de enlace - logit)
pred_lin_sim_pi=cbind(rep(1,length(BBDD$urbano)),BBDD$urbano,
BBDD$pared_cemento,BBDD$piso_tierra,BBDD$vivienda_propia,
BBDD$gasto_agua,BBDD$niv_edu_sup,BBDD$percepcion_muy_mala,
BBDD$estr_social_alto,BBDD$alquiler,BBDD$provincia_1102,
BBDD$provincia_1103,BBDD$provincia_1104,
BBDD$provincia_1105)%*%beta_est+gamma_sim_x
pi_hat_sim=as.numeric(exp(pred_lin_sim_pi)/(1+exp(pred_lin_sim_pi)))
pi_hat_area_sim=aggregate(pi_hat_sim,list(BBDD$ubigeo), FUN=mean)$x

#Simulación nueva de Y poblacional según pi_hat_sim y estimación del valor esperado de H
Y_sim_muestra=unlist(lapply(pi_hat_sim,rbinom,n=1,size=1))

Y_sim=NULL
datos_sim_ubigeo=NULL

for (area in 1:J) {
ubigeo_sim=as.numeric(DatosCenso2017$ubigeo[area])
Y_sim=rbind(Y_sim,as.matrix(unlist(mapply(rbinom,
prob=pi_hat_sim[BBDD$ubigeo==ubigeo_sim],
n=rep(tasa_rep_nj[area],times=nj[area]),
size=rep(1,nj[area]),
SIMPLIFY = FALSE))))

datos_sim_ubigeo=rbind(datos_sim_ubigeo,
as.matrix(rep(ubigeo_sim,times=tasa_rep_nj[area]*nj[area])))
```

```
}

Esim_Hj=fmean(data.frame(Y_sim,datos_sim_ubigeo),datos_sim_ubigeo)$Y_sim
Esim_Hj_prov=fmean(data.frame(Y_sim,floor(datos_sim_ubigeo/100)),
floor(datos_sim_ubigeo/100))$Y_sim
Esim_Hj_depa=mean(Y_sim)

#Estimación nueva de GAMLSS con datos simulados de Y
muestra_boot=data.frame(saneamiento=Y_sim_muestra,ubigeo=BBDD$ubigeo,
urbano=BBDD$urbano,pared_cemento=BBDD$pared_cemento,
piso_tierra=BBDD$piso_tierra,vivienda_propia=BBDD$vivienda_propia,
gasto_agua=BBDD$gasto_agua,niv_edu_sup=BBDD$niv_edu_sup,
percepcion_muy_mala=BBDD$percepcion_muy_mala,
estr_social_alto=BBDD$estr_social_alto,
alquiler=BBDD$alquiler,provincia_1102=BBDD$provincia_1102,
provincia_1103=BBDD$provincia_1103,provincia_1104=BBDD$provincia_1104,
provincia_1105=BBDD$provincia_1105,factor07=BBDD$factor07)

modelo_gamlss_boot=gamlss(saneamiento~urbano+pared_cemento+piso_tierra+
vivienda_propia+gasto_agua+niv_edu_sup+
percepcion_muy_mala+estr_social_alto+
provincia_1102+provincia_1103+provincia_1104+provincia_1105+
alquiler+re(random=~1|ubigeo),
family = BI(mu.link="logit"),data = muestra_boot,weights = factor07,
control = gamlss.control(n.cyc = 40,c.crit = 0.01))

#Nueva estimación de pi_hat
pi_hat_boot=fitted(modelo_gamlss_boot)
names(pi_hat_boot)=BBDD$ubigeo

#Estimación de Hj_sim
Hj_simboot=NULL
Hj_simboot_prov=NULL
Hj_simboot_depa=NULL

for (k in 1:L) {

datos_sim_Y=mapply(rbinom,prob=pi_hat_boot,n=rep(tasa_rep_nj_ajust,times=nj),
SIMPLIFY=TRUE,size=1)

id_area_faltantes=as.numeric(rep(names(datos_sim_Y), lengths(datos_sim_Y)))
datos_sim_Y_faltantes=unlist(datos_sim_Y,use.names = FALSE)
datos_sim_faltantes=data.frame(datos_sim_Y=datos_sim_Y_faltantes,
id_area=id_area_faltantes)
datos_sim_final=rbind(data.frame(datos_sim_Y=BBDD$saneamiento,
id_area=BBDD$ubigeo),datos_sim_faltantes)
```

```
datos_sim_final$prov=floor(datos_sim_final$id_area/100)

Hj_sim=fmean(datos_sim_final[,c(1,2)],datos_sim_final$id_area)$datos_sim_Y
Hj_simboot=cbind(Hj_simboot,Hj_sim)

Hj_sim_prov=fmean(datos_sim_final[,c(1,3)],datos_sim_final$prov)$datos_sim_Y
Hj_simboot_prov=cbind(Hj_simboot_prov,Hj_sim_prov)

Hj_sim_depa=mean(datos_sim_final$datos_sim_Y)
Hj_simboot_depa=cbind(Hj_simboot_depa,Hj_sim_depa)

}

est_gamlss_boot=as.numeric(rowMeans(Hj_simboot))
desv_cuadra_boot=cbind(desv_cuadra_boot,(est_gamlss_boot-Esim_Hj)^2)
dist_gamlss_boot=cbind(dist_gamlss_boot,est_gamlss_boot)

est_gamlss_boot_prov=as.numeric(rowMeans(Hj_simboot_prov))
desv_cuadra_boot_prov=cbind(desv_cuadra_boot_prov,
(est_gamlss_boot_prov-Esim_Hj_prov)^2)
dist_gamlss_boot_prov=cbind(dist_gamlss_boot_prov,
est_gamlss_boot_prov)

est_gamlss_boot_depa=as.numeric(rowMeans(Hj_simboot_depa))
desv_cuadra_boot_depa=cbind(desv_cuadra_boot_depa,
(est_gamlss_boot_depa-Esim_Hj_depa)^2)
dist_gamlss_boot_depa=cbind(dist_gamlss_boot_depa,est_gamlss_boot_depa)

print(b)

}

MSE_gamlss_boot=as.numeric(rowMeans(desv_cuadra_boot)) ; mean(MSE_gamlss_boot)
MSE_gamlss_boot_prov=as.numeric(rowMeans(desv_cuadra_boot_prov)) ; mean(MSE_gamlss_boot_prov)
MSE_gamlss_boot_depa=as.numeric(rowMeans(desv_cuadra_boot_depa)) ; mean(MSE_gamlss_boot_depa)

t_new=Sys.time() - t_old
print(t_new)

IC_sup_gamlss=est_gamlss+1.96*sqrt(MSE_gamlss_boot)
IC_sup_gamlss[IC_sup_gamlss>1]=1
IC_inf_gamlss=est_gamlss-1.96*sqrt(MSE_gamlss_boot)
IC_inf_gamlss[IC_inf_gamlss<0]=0

IC_sup_gamlss_prov=est_gamlss_prov+1.96*sqrt(MSE_gamlss_boot_prov)
IC_sup_gamlss_prov[IC_sup_gamlss_prov>1]=1
```

```
IC_inf_gamlss_prov=est_gamlss_prov-1.96*sqrt(MSE_gamlss_boot_prov)
IC_inf_gamlss_prov[IC_inf_gamlss_prov<0]=0

IC_sup_gamlss_depa=est_gamlss_depa+1.96*sqrt(MSE_gamlss_boot_depa)
IC_sup_gamlss_depa[IC_sup_gamlss_depa>1]=1
IC_inf_gamlss_depa=est_gamlss_depa-1.96*sqrt(MSE_gamlss_boot_depa)
IC_inf_gamlss_depa[IC_inf_gamlss_depa<0]=0

#contraste de resultados con medias reales
Diff_gamlss=abs(est_gamlss-prom_Y_area)/prom_Y_area ; mean(Diff_gamlss)
NoDentroIC_gamlss=prom_Y_area<IC_inf_gamlss | prom_Y_area>IC_sup_gamlss
mean(NoDentroIC_gamlss)

Diff_gamlss_prov=abs(est_gamlss_prov-prom_Y_area_prov)/prom_Y_area_prov
mean(Diff_gamlss_prov)
NoDentroIC_gamlss_prov=prom_Y_area_prov<IC_inf_gamlss_prov | prom_Y_area_prov>IC_sup_gamlss_prov
mean(NoDentroIC_gamlss_prov)

Diff_gamlss_depa=abs(est_gamlss_depa-prom_Y_area_depa)/prom_Y_area_depa
mean(Diff_gamlss_depa)
NoDentroIC_gamlss_depa=prom_Y_area_depa<IC_inf_gamlss_depa | prom_Y_area_depa>IC_sup_gamlss_depa
mean(NoDentroIC_gamlss_depa)

######TABLA DE INDICADORES######
tabla_indicadores_dist=data.frame(ubigeo=DatosCenso2017$ubigeo,
n=nj,H=est_gamlss,IC_inf=IC_inf_gamlss,IC_sup=IC_sup_gamlss)
tabla_indicadores_prov=data.frame(ubigeo=DatosCenso2017_prov$ubigeo,
n=nj_prov,H=est_gamlss_prov,IC_inf=IC_inf_gamlss_prov,
IC_sup=IC_sup_gamlss_prov)
tabla_indicadores_depa=data.frame(ubigeo="11",n=nj_depa,H=est_gamlss_depa,
IC_inf=IC_inf_gamlss_depa,IC_sup=IC_sup_gamlss_depa)

rbind(tabla_indicadores_depa,tabla_indicadores_prov,tabla_indicadores_dist)

######RESUMEN DE RESULTADOS AGREGADOS####
tabla=rbind(cbind(mean(MSE_directo),mean(MSE_gamlss_boot)),
cbind(mean(Diff_directo),mean(Diff_gamlss)),
cbind(mean(NoDentroIC_directo),mean(NoDentroIC_gamlss)))

rownames(tabla)=c("MSE","Diff. Valor real","% Áreas no cubiertas en IC")
colnames(tabla)=c("Est. Directa","Est. GAMLSS")
tabla

tabla_prov=rbind(cbind(mean(MSE_directo_prov),mean(MSE_gamlss_boot_prov)),
cbind(mean(Diff_directo_prov),mean(Diff_gamlss_prov)),
cbind(mean(NoDentroIC_directo_prov),mean(NoDentroIC_gamlss_prov)))
```

```
rownames(tabla_prov)=c("MSE","Diff. Valor real","% Áreas no cubiertas en IC")
colnames(tabla_prov)=c("Est. Directa","Est. GAMLSS")
tabla_prov

tabla_depa=rbind(cbind(mean(MSE_directo_depa),mean(MSE_gamlss_boot_depa)),
cbind(mean(Diff_directo_depa),mean(Diff_gamlss_depa)),
cbind(mean(NoDentroIC_directo_depa),mean(NoDentroIC_gamlss_depa)))

rownames(tabla_depa)=c("MSE","Diff. Valor real","% Áreas no cubiertas en IC")
colnames(tabla_depa)=c("Est. Directa","Est. GAMLSS")
tabla_depa

#####RESUMEN DE RESULTADOS POR TAMAÑO DE MUESTRA####

#MSE
datos_MSE_nj=data.frame(nj=c(nj,nj_prov,nj_depa),
MSE_directo=c(MSE_directo,MSE_directo_prov,MSE_directo_depa),
MSE_gamlss_boot=c(MSE_gamlss_boot,MSE_gamlss_boot_prov,
MSE_gamlss_boot_depa))
plot(datos_MSE_nj$nj,datos_MSE_nj$MSE_directo,
main=NULL,
xlab="Tamaños de muestra",
ylab="MSE",
pch=19,col="salmon",xlim = c(0,800))
points(datos_MSE_nj$nj,datos_MSE_nj$MSE_gamlss_boot,
pch=19,col="lightblue")
legend("topright", c("Est. Directa","Est. GAMLSS"),pch=19,
col=c("salmon","lightblue"),bty="n",cex=1)

#Diff Valor real
datos_diff_nj=data.frame(nj=c(nj,nj_prov,nj_depa),
Diff_directo=c(Diff_directo,Diff_directo_prov,Diff_directo_depa),
Diff_gamlss=c(Diff_gamlss,Diff_gamlss_prov,Diff_gamlss_depa))
plot(datos_diff_nj$nj,datos_diff_nj$Diff_directo,
main=NULL,
xlab="Tamaños de muestra",
ylab="Diferencia relativa con valor real",
pch=19,col="salmon",xlim = c(0,800))
points(datos_diff_nj$nj,datos_diff_nj$Diff_gamlss,
pch=19,col="lightblue")
legend("topright", c("Est. Directa","Est. GAMLSS"),pch=19,
col=c("salmon","lightblue"),bty="n",cex=1)

nj_intervalos=cut(c(nj,nj_prov,nj_depa),breaks = c(0,10,20,50,100,2000),
labels = c("Menor a 10","Entre 10 y 20","Entre 20 y 50","Entre 50 y 100","Más de 100"))
```

```
datos_est_nj=data.frame(nj_intervalos,
niv_geo=c(rep("Distrito",37),rep("Provincia",5),"Departamento"),
est_indicador=c(est_indicador,est_indicador_prov,est_indicador_depa),
est_gamlss=c(est_gamlss,est_gamlss_prov,est_gamlss_depa),
prom_Y_area=c(prom_Y_area,prom_Y_area_prov,prom_Y_area_depa))

ggplot(datos_est_nj, aes(x=est_indicador, y=prom_Y_area, color=nj_intervalos,shape=niv_geo)) +
geom_point(size=4,alpha = 7/10)+ geom_abline(slope=1, intercept=0, col = "red") +
xlab("Estimaciones directas") + ylab("Valores poblacionales") + ylim(0, 1) + xlim(0,1)+
labs(color = "Tamaños de muestra",shape="División geográfica") +
theme(legend.position = "none")

ggplot(datos_est_nj, aes(x=est_gamlss, y=prom_Y_area, color=nj_intervalos,shape=niv_geo)) +
geom_point(size=4,alpha = 7/10)+ geom_abline(slope=1, intercept=0, col = "red") +
xlab("Estimaciones GAMLSS") + ylab("Valores poblacionales")+ ylim(0, 1) + xlim(0,1)+
labs(color = "Tamaños de muestra",shape="División geográfica")

#Grafico IC - GAMLSS (departamento, provincias y distritos)
plot(c(est_gamlss_depa,est_gamlss_prov,est_gamlss),pch=19,main = NULL,
xlab = "",ylab = "Valor estimado",ylim = c(0,1),xaxt = "n")
arrows(1:(1+5+J), c(IC_inf_gamlss_depa,IC_inf_gamlss_prov,IC_inf_gamlss),
1:(1+5+J), c(IC_sup_gamlss_depa,IC_sup_gamlss_prov,IC_sup_gamlss),
length=0.05, angle=90, code=3)
points(c(prom_Y_area_depa,prom_Y_area_prov,prom_Y_area),
col=rgb(red = 1, green = 0, blue = 0,alpha=0.6),pch=19)
legend("bottomright", c("Est. GAMLSS","Valor real"),pch=19,
col=c("black","firebrick1"),bty="n",cex=0.8)
axis(1, at=1:(1+5+J), labels=c("11",
DatosCenso2017_prov$ubigeo,
DatosCenso2017$ubigeo),las=3,cex.axis=0.8)
```

7.3. Códigos de R de GAMLSS para la estimación de indicadores de cantidad de habitaciones de las viviendas

```
library(haven)
library(readxl)
library(gamlss.dist)
library(gamlss.ggplots)
library(gamlss)
library(sae)
library(dplyr)
library(sampling)
library(survey)
library(ggplot2)
```

```
library(collapse)
library(gridExtra)

####CARGA DE DATOS#####

setwd("C:/Users/hans_/Desktop/maestria/PUCP/Tesis 2")

load("BBDDENAHO2017.Rda")
DatosCenso2017=read_excel("DatosCenso2017.xlsx",
col_types = c("text", "numeric", "numeric","numeric"),
sheet = "Ica")

DatosCenso2017_prov=read_excel("DatosCenso2017.xlsx",
col_types = c("text", "numeric", "numeric","numeric"),
sheet = "Ica_prov")
DatosCenso2017_depa=read_excel("DatosCenso2017.xlsx",
col_types = c("text", "numeric", "numeric","numeric"),
sheet = "Ica_depa")

prom_Y_area=DatosCenso2017$habitaciones
prom_Y_area_prov=DatosCenso2017_prov$habitaciones
prom_Y_area_depa=DatosCenso2017_depa$habitaciones

Nj=DatosCenso2017$Nj
Nj_prov=DatosCenso2017_prov$Nj
Nj_depa=DatosCenso2017_depa$Nj

ubigeo=rep(as.numeric(DatosCenso2017$ubigeo),times=Nj)
J=length(as.numeric(DatosCenso2017$ubigeo))

BBDDENAHO2017$provincia=substr(BBDDENAHO2017$ubigeo,1,4)

BBDD=BBDDENAHO2017[complete.cases(BBDDENAHO2017),]

#incorporacion de dummies de provincia
BBDD$provincia_1102=BBDD$provincia=="1102"
BBDD$provincia_1103=BBDD$provincia=="1103"
BBDD$provincia_1104=BBDD$provincia=="1104"
BBDD$provincia_1105=BBDD$provincia=="1105"

#incorporacion de término de iteraccion entre alquiler y gasto pagado
BBDD$alquiler_gasto=BBDD$alquiler*BBDD$gasto_pagado

barplot(table(BBDD$habitaciones))

#NB GLM se usa para identificar valores extremos
```

```
modelo_glm_nb=glm.nb(habitaciones~alquiler+gasto_pagado+alquiler_gasto+m_hogar+vivienda_propia+
piso_tierra+niv_edu_sup+estr_social_alto+percepcion_muy_mala+
provincia_1102+provincia_1103+provincia_1104+provincia_1105,
weights=factor07,data = BBDD,link=log)
summary(modelo_glm_nb)

plot(cooks.distance(modelo_glm_nb))
BBDD=BBDD[cooks.distance(modelo_glm_nb)<1,]

nj=as.numeric(table(BBDD$ubigeo))
nj_prov=as.numeric(table(BBDD$provincia))
nj_depa=length(BBDD$provincia)

tasa_rep_nj=ceiling(Nj/nj)

hist(prom_Y_area,main = NULL,
xlab="Distribución de frecuencias de la cantidad promedio de habitaciones por vivienda",
ylab="Frencuencias") ; summary(prom_Y_area)
boxplot(nj,xlab="Tamaños de muestra de los distritos de Ica",horizontal = TRUE)
summary(nj)

#####ESTIMACION DIRECTA CON PROMEDIOS#######
BBDD$ubigeo=as.numeric(BBDD$ubigeo)

#distrital
diseño_muestral=svydesign(ids=~conglome,weights=~factor07,strata=~estrato,data = BBDD)
est_indicador=svyby(~habitaciones,~ubigeo,diseño_muestral,svymean)$habitaciones
var_indicador=(svyby(~habitaciones,~ubigeo,diseño_muestral,svymean)$se)^2
MSE_directo=var_indicador

#IC
IC_sup_directo=est_indicador+1.96*sqrt(var_indicador)
IC_inf_directo=est_indicador-1.96*sqrt(var_indicador)

#contraste de resultados con medias reales
Diff_directo=abs(est_indicador-prom_Y_area)/prom_Y_area ; mean(Diff_directo)
NoDentroIC_directo=prom_Y_area<IC_inf_directo | prom_Y_area>IC_sup_directo
mean(NoDentroIC_directo)

#provincial
est_indicador_prov=svyby(~habitaciones,~provincia,diseño_muestral,svymean)$habitaciones
var_indicador_prov=(svyby(~habitaciones,~provincia,diseño_muestral,svymean)$se)^2
MSE_directo_prov=var_indicador_prov

#IC
IC_sup_directo_prov=est_indicador_prov+1.96*sqrt(var_indicador_prov)
```

```
IC_inf_directo_prov=est_indicador_prov-1.96*sqrt(var_indicador_prov)

#contraste de resultados con medias reales
Diff_directo_prov=abs(est_indicador_prov-prom_Y_area_prov)/prom_Y_area_prov ; mean(Diff_directo_prov)
NoDentroIC_directo_prov=prom_Y_area_prov<IC_inf_directo_prov | prom_Y_area_prov>IC_sup_directo_prov
mean(NoDentroIC_directo_prov)

#departamental
est_indicador_depa=coef(svymean(~habitaciones,diseño_muestral))
var_indicador_depa=SE(svymean(~habitaciones,diseño_muestral))^2
MSE_directo_depa=var_indicador_depa

#IC
IC_sup_directo_depa=est_indicador_depa+1.96*sqrt(var_indicador_depa)
IC_inf_directo_depa=est_indicador_depa-1.96*sqrt(var_indicador_depa)

#contraste de resultados con medias reales
Diff_directo_depa=abs(est_indicador_depa-prom_Y_area_depa)/prom_Y_area_depa
mean(Diff_directo_depa)
NoDentroIC_directo_depa=prom_Y_area_depa<IC_inf_directo_depa | prom_Y_area_depa>IC_sup_directo_depa
mean(NoDentroIC_directo_depa)

#####ESTIMACION GAMLSS#####
t_old=Sys.time()

set.seed(432)

#Modelización del parámetro de la distribución de la variable discreta
modelo_gamlss=gamlss(habitaciones~alquiler+gasto_pagado+alquiler_gasto+m_hogar+
vivienda_propia+piso_tierra+niv_edu_sup+estr_social_alto+percepcion_muy_mala+
provincia_1102+provincia_1103+provincia_1104+provincia_1105+
re(random=~1|ubigeo),
sigma.formula=habitaciones~gasto_pagado+alquiler+estr_social_alto+
re(random=~1|ubigeo),
family = NBI(mu.link = "log", sigma.link ="log"),
control = gamlss.control(n.cyc = 40,c.crit = 0.01,sigma.step = 0.1),
sigma.start = 0.000000000001,weights=factor07,
data = BBDD)

summary(modelo_gamlss)
plot(modelo_gamlss)

grid.arrange(
resid_mu(modelo_gamlss,title="Residuals vs Fitted Values",value = 3,annotate = FALSE),
resid_index(modelo_gamlss,title="Residuals vs Index",value = 3,annotate = FALSE),
resid_density(modelo_gamlss,title="Density Estimate"),
```

```
resid_qqplot(modelo_gamlss, value = 3, points.col = "steelblue4",
line.col = "darkgray", check_overlap = TRUE, title="Q-Q Plot"))

#Test de normalidad de errores (randomized-quantile residuales)
shapiro.test(resid(modelo_gamlss))
ks.test(resid(modelo_gamlss),"pnorm")

mu_hat=fitted(modelo_gamlss,what = "mu")
#valores estimados del parámetro de la distribución negative binomial

sigma_hat=fitted(modelo_gamlss,what = "sigma")
#valores estimados del parámetro de la distribución  negative binomial

names(mu_hat)=BBDD$ubigeo
names(sigma_hat)=BBDD$ubigeo

#Estimación de indicador - según método de Graf, Marin y Molina (2019)

#Simulacion de datos sintéticos de Y y estimación de indicador
L=200

datos_faltantes=Nj-nj
tasa_rep_nj_ajust=datos_faltantes/nj
Hj_simMC=NULL
Hj_simMC_prov=NULL
Hj_simMC_depa=NULL

for (k in 1:L) {

datos_sim_Y=mapply(rNBI,mu=mu_hat,sigma=sigma_hat,n=rep(tasa_rep_nj_ajust,times=nj),
SIMPLIFY=TRUE)

id_area_faltantes=as.numeric(rep(names(datos_sim_Y), lengths(datos_sim_Y)))
datos_sim_Y_faltantes=unlist(datos_sim_Y,use.names = FALSE)
datos_sim_faltantes=data.frame(datos_sim_Y=datos_sim_Y_faltantes,
id_area=id_area_faltantes)
datos_sim_final=rbind(data.frame(datos_sim_Y=BBDD$habitaciones,
id_area=BBDD$ubigeo),datos_sim_faltantes)
datos_sim_final$prov=floor(datos_sim_final$id_area/100)

Hj_sim=fmean(datos_sim_final[,c(1,2)],datos_sim_final$id_area)$datos_sim_Y
Hj_simMC=cbind(Hj_simMC,Hj_sim)

Hj_sim_prov=fmean(datos_sim_final[,c(1,3)],datos_sim_final$prov)$datos_sim_Y
Hj_simMC_prov=cbind(Hj_simMC_prov,Hj_sim_prov)
```

```
Hj_sim_depa=mean(datos_sim_final$datos_sim_Y)
Hj_simMC_depa=cbind(Hj_simMC_depa,Hj_sim_depa)
}

est_gamlss=as.numeric(rowMeans(Hj_simMC))
est_gamlss_prov=as.numeric(rowMeans(Hj_simMC_prov))
est_gamlss_depa=as.numeric(rowMeans(Hj_simMC_depa))

#Estimacion de MSE en GAMLSS

set.seed(432)
B=200

beta_est_mu=modelo_gamlss$mu.coefficients[1:14] # coeficientes mu
beta_est_sigma=modelo_gamlss$sigma.coefficients[1:4] # coeficientes sigma

# efectos aleatorios
gamma_est_mu=as.numeric(modelo_gamlss$mu.coefSmo[[1]]$coefficients$random$ubigeo)

# efectos aleatorios
gamma_est_sigma=as.numeric(modelo_gamlss$sigma.coefSmo[[1]]$coefficients$random$ubigeo)

var_gamma_mu=as.numeric(VarCorr(modelo_gamlss$mu.coefSmo[[1]])[1])
m_var_cov_mu=diag(var_gamma_mu,J)

var_gamma_sigma=as.numeric(VarCorr(modelo_gamlss$sigma.coefSmo[[1]])[1])
m_var_cov_sigma=diag(var_gamma_sigma,J)

desv_cuadra_boot=NULL
dist_gamlss_boot=NULL

desv_cuadra_boot_prov=NULL
dist_gamlss_boot_prov=NULL

desv_cuadra_boot_depa=NULL
dist_gamlss_boot_depa=NULL

for (b in 1:B) {

#Simulación nueva de gamma
t_sim_mu=rnorm(J)
gamma_sim_mu=sqrt(m_var_cov_mu)%*%t_sim_mu
gamma_sim_x_mu=rep(gamma_sim_mu,times=nj)

t_sim_sigma=rnorm(J)
gamma_sim_sigma=sqrt(m_var_cov_sigma)%*%t_sim_sigma
```

```
gamma_sim_x_sigma=rep(gamma_sim_sigma,times=nj)

#Simulación nueva de mu_hat y sigma_hat (inversa de función de enlace - log)
pred_lin_sim_mu=cbind(rep(1,length(BBDD$habitaciones)),
BBDD$alquiler,BBDD$gasto_pagado,BBDD$alquiler_gasto,BBDD$m_hogar,
BBDD$vivienda_propia,BBDD$piso_tierra,BBDD$niv_edu_sup,
BBDD$estr_social_alto,
BBDD$percepcion_muy_mala,BBDD$provincia_1102,
BBDD$provincia_1103,BBDD$provincia_1104,
BBDD$provincia_1105)%*%beta_est_mu+gamma_sim_x_mu
pred_lin_sim_sigma=cbind(rep(1,length(BBDD$habitaciones)),
BBDD$gasto_pagado,BBDD$alquiler,
BBDD$estr_social_alto)%*%beta_est_sigma+gamma_sim_x_sigma

mu_sim=exp(pred_lin_sim_mu)
sigma_sim=exp(pred_lin_sim_sigma)

#Simulación nueva de Y poblacional según mu_hat_sim y sigma_hat_sim
#y estimación del valor esperado de H
Y_sim_muestra=unlist(mapply(rNBI,mu=mu_sim,sigma=sigma_sim,n=rep(1,length(mu_sim)),
SIMPLIFY=TRUE))

Y_sim=NULL
datos_sim_ubigeo=NULL

for (area in 1:J) {
ubigeo_sim=as.numeric(DatosCenso2017$ubigeo[area])
Y_sim=rbind(Y_sim,as.matrix(unlist(mapply(rNBI,
mu=mu_sim[BBDD$ubigeo==ubigeo_sim],
sigma=sigma_sim[BBDD$ubigeo==ubigeo_sim],
n=rep(tasa_rep_nj[area],times=nj[area]),
SIMPLIFY = FALSE))))

datos_sim_ubigeo=rbind(datos_sim_ubigeo,
as.matrix(rep(ubigeo_sim,times=tasa_rep_nj[area]*nj[area])))
}

Esim_Hj=fmean(data.frame(Y_sim,datos_sim_ubigeo),datos_sim_ubigeo)$Y_sim
Esim_Hj_prov=fmean(data.frame(Y_sim,floor(datos_sim_ubigeo/100)),
floor(datos_sim_ubigeo/100))$Y_sim
Esim_Hj_depa=mean(Y_sim)

#Estimación nueva de GAMLSS con datos simulados de Y
muestra_boot=data.frame(habitaciones=Y_sim_muestra,ubigeo=BBDD$ubigeo,
alquiler=BBDD$alquiler,gasto_pagado=BBDD$gasto_pagado,
alquiler_gasto=BBDD$alquiler_gasto,
```

```
m_hogar=BBDD$m_hogar,vivienda_propia=BBDD$vivienda_propia,
piso_tierra=BBDD$piso_tierra,niv_edu_sup=BBDD$niv_edu_sup,
estr_social_alto=BBDD$estr_social_alto,provincia_1102=BBDD$provincia_1102,
provincia_1103=BBDD$provincia_1103,provincia_1104=BBDD$provincia_1104,
provincia_1105=BBDD$provincia_1105,
percepcion_muy_mala=BBDD$percepcion_muy_mala,factor07=BBDD$factor07)
modelo_gamlss_boot=gamlss(habitaciones~alquiler+gasto_pagado+alquiler_gasto+m_hogar+vivienda_propia+
piso_tierra+niv_edu_sup+estr_social_alto+
percepcion_muy_mala+provincia_1102+provincia_1103+
provincia_1104+provincia_1105+
re(random=~1|ubigeo),
sigma.formula=habitaciones~gasto_pagado+alquiler+
estr_social_alto+re(random=~1|ubigeo),
family = NBI(mu.link = "log", sigma.link ="log"),
control = gamlss.control(n.cyc = 40,c.crit = 0.01,sigma.step = 0.1),
sigma.start = 0.000000000001,weights=factor07,
data = muestra_boot)

#Nueva estimación de mu y sigma
mu_boot=fitted(modelo_gamlss_boot,what = "mu")
sigma_boot=fitted(modelo_gamlss_boot,what = "sigma")

names(mu_boot)=BBDD$ubigeo
names(sigma_boot)=BBDD$ubigeo

#Estimación de Hj_sim
Hj_simboot=NULL
Hj_simboot_prov=NULL
Hj_simboot_depa=NULL

for (k in 1:L) {

datos_sim_Y=mapply(rNBI,mu=mu_boot,sigma=sigma_boot,
n=rep(tasa_rep_nj_ajust,times=nj),
SIMPLIFY=TRUE)

id_area_faltantes=as.numeric(rep(names(datos_sim_Y), lengths(datos_sim_Y)))
datos_sim_Y_faltantes=unlist(datos_sim_Y,use.names = FALSE)
datos_sim_faltantes=data.frame(datos_sim_Y=datos_sim_Y_faltantes,
id_area=id_area_faltantes)
datos_sim_final=rbind(data.frame(datos_sim_Y=BBDD$habitaciones,
id_area=BBDD$ubigeo),datos_sim_faltantes)
datos_sim_final$prov=floor(datos_sim_final$id_area/100)

Hj_sim=fmean(datos_sim_final[,c(1,2)],datos_sim_final$id_area)$datos_sim_Y
Hj_simboot=cbind(Hj_simboot,Hj_sim)
```

```
Hj_sim_prov=fmean(datos_sim_final[,c(1,3)],datos_sim_final$prov)$datos_sim_Y
Hj_simboot_prov=cbind(Hj_simboot_prov,Hj_sim_prov)

Hj_sim_depa=mean(datos_sim_final$datos_sim_Y)
Hj_simboot_depa=cbind(Hj_simboot_depa,Hj_sim_depa)

}

est_gamlss_boot=as.numeric(rowMeans(Hj_simboot))
desv_cuadra_boot=cbind(desv_cuadra_boot,(est_gamlss_boot-Esim_Hj)^2)
dist_gamlss_boot=cbind(dist_gamlss_boot,est_gamlss_boot)

est_gamlss_boot_prov=as.numeric(rowMeans(Hj_simboot_prov))
desv_cuadra_boot_prov=cbind(desv_cuadra_boot_prov,(est_gamlss_boot_prov-Esim_Hj_prov)^2)
dist_gamlss_boot_prov=cbind(dist_gamlss_boot_prov,est_gamlss_boot_prov)

est_gamlss_boot_depa=as.numeric(rowMeans(Hj_simboot_depa))
desv_cuadra_boot_depa=cbind(desv_cuadra_boot_depa,(est_gamlss_boot_depa-Esim_Hj_depa)^2)
dist_gamlss_boot_depa=cbind(dist_gamlss_boot_depa,est_gamlss_boot_depa)

print(b)
}

MSE_gamlss_boot=as.numeric(rowMeans(desv_cuadra_boot))
mean(MSE_gamlss_boot)
MSE_gamlss_boot_prov=as.numeric(rowMeans(desv_cuadra_boot_prov))
mean(MSE_gamlss_boot_prov)
MSE_gamlss_boot_depa=as.numeric(rowMeans(desv_cuadra_boot_depa))
mean(MSE_gamlss_boot_depa)

t_new=Sys.time() - t_old
print(t_new)

IC_sup_gamlss=est_gamlss+1.96*sqrt(MSE_gamlss_boot)
IC_inf_gamlss=est_gamlss-1.96*sqrt(MSE_gamlss_boot)

IC_sup_gamlss_prov=est_gamlss_prov+1.96*sqrt(MSE_gamlss_boot_prov)
IC_inf_gamlss_prov=est_gamlss_prov-1.96*sqrt(MSE_gamlss_boot_prov)

IC_sup_gamlss_depa=est_gamlss_depa+1.96*sqrt(MSE_gamlss_boot_depa)
IC_inf_gamlss_depa=est_gamlss_depa-1.96*sqrt(MSE_gamlss_boot_depa)

#contraste de resultados con medias reales
Diff_gamlss=abs(est_gamlss-prom_Y_area)/prom_Y_area ; mean(Diff_gamlss)
NoDentroIC_gamlss=prom_Y_area<IC_inf_gamlss | prom_Y_area>IC_sup_gamlss
```

```
mean(NoDentroIC_gamlss)

Diff_gamlss_prov=abs(est_gamlss_prov-prom_Y_area_prov)/prom_Y_area_prov
mean(Diff_gamlss_prov)
NoDentroIC_gamlss_prov=prom_Y_area_prov<IC_inf_gamlss_prov | prom_Y_area_prov>IC_sup_gamlss_prov
mean(NoDentroIC_gamlss_prov)

Diff_gamlss_depa=abs(est_gamlss_depa-prom_Y_area_depa)/prom_Y_area_depa
mean(Diff_gamlss_depa)
NoDentroIC_gamlss_depa=prom_Y_area_depa<IC_inf_gamlss_depa | prom_Y_area_depa>IC_sup_gamlss_depa
mean(NoDentroIC_gamlss_depa)

######TABLA DE INDICADORES######
tabla_indicadores_dist=data.frame(ubigeo=DatosCenso2017$ubigeo,
n=nj,H=est_gamlss,IC_inf=IC_inf_gamlss,
IC_sup=IC_sup_gamlss)
tabla_indicadores_prov=data.frame(ubigeo=DatosCenso2017_prov$ubigeo,n=nj_prov,
H=est_gamlss_prov,IC_inf=IC_inf_gamlss_prov,
IC_sup=IC_sup_gamlss_prov)
tabla_indicadores_depa=data.frame(ubigeo="11",n=nj_depa,H=est_gamlss_depa,
IC_inf=IC_inf_gamlss_depa,IC_sup=IC_sup_gamlss_depa)

rbind(tabla_indicadores_depa,tabla_indicadores_prov,tabla_indicadores_dist)

######RESUMEN DE RESULTADOS AGREGADOS####
tabla=rbind(cbind(mean(MSE_directo),mean(MSE_gamlss_boot)),
cbind(mean(Diff_directo),mean(Diff_gamlss)),
cbind(mean(NoDentroIC_directo),mean(NoDentroIC_gamlss)))

rownames(tabla)=c("MSE","Diff. Valor real","% Áreas no cubiertas en IC")
colnames(tabla)=c("Est. Directa","Est. GAMLSS")
tabla

tabla_prov=rbind(cbind(mean(MSE_directo_prov),mean(MSE_gamlss_boot_prov)),
cbind(mean(Diff_directo_prov),mean(Diff_gamlss_prov)),
cbind(mean(NoDentroIC_directo_prov),mean(NoDentroIC_gamlss_prov)))

rownames(tabla_prov)=c("MSE","Diff. Valor real","% Áreas no cubiertas en IC")
colnames(tabla_prov)=c("Est. Directa","Est. GAMLSS")
tabla_prov

tabla_depa=rbind(cbind(mean(MSE_directo_depa),mean(MSE_gamlss_boot_depa)),
cbind(mean(Diff_directo_depa),mean(Diff_gamlss_depa)),
cbind(mean(NoDentroIC_directo_depa),mean(NoDentroIC_gamlss_depa)))

rownames(tabla_depa)=c("MSE","Diff. Valor real","% Áreas no cubiertas en IC")
```

```
colnames(tabla_depa)=c("Est. Directa","Est. GAMLSS")
tabla_depa

#####RESUMEN DE RESULTADOS POR TAMAÑO DE MUESTRA####

#MSE
datos_MSE_nj=data.frame(nj=c(nj,nj_prov,nj_depa),
MSE_directo=c(MSE_directo,MSE_directo_prov,MSE_directo_depa),
MSE_gamlss_boot=c(MSE_gamlss_boot,MSE_gamlss_boot_prov,MSE_gamlss_boot_depa))
plot(datos_MSE_nj$nj,datos_MSE_nj$MSE_directo,
main=NULL,
xlab="Tamaños de muestra",
ylab="MSE",
pch=19,col="salmon",xlim = c(0,800))
points(datos_MSE_nj$nj,datos_MSE_nj$MSE_gamlss_boot,
pch=19,col="lightblue")
legend("topright", c("Est. Directa","Est. GAMLSS"),pch=19,
col=c("salmon","lightblue"),bty="n",cex=1)

#Diff Valor real
datos_diff_nj=data.frame(nj=c(nj,nj_prov,nj_depa),
Diff_directo=c(Diff_directo,Diff_directo_prov,Diff_directo_depa),
Diff_gamlss=c(Diff_gamlss,Diff_gamlss_prov,Diff_gamlss_depa))
plot(datos_diff_nj$nj,datos_diff_nj$Diff_directo,
main=NULL,
xlab="Tamaños de muestra",
ylab="Diferencia relativa con valor real",
pch=19,col="salmon",xlim = c(0,800))
points(datos_diff_nj$nj,datos_diff_nj$Diff_gamlss,
pch=19,col="lightblue")
legend("topright", c("Est. Directa","Est. GAMLSS"),pch=19,
col=c("salmon","lightblue"),bty="n",cex=1)

nj_intervalos=cut(c(nj,nj_prov,nj_depa),breaks = c(0,10,20,50,100,2000),
labels = c("Menor a 10","Entre 10 y 20","Entre 20 y 50","Entre 50 y 100","Más de 100"))

datos_est_nj=data.frame(nj_intervalos,
niv_geo=c(rep("Distrito",37),rep("Provincia",5),"Departamento"),
est_indicador=c(est_indicador,est_indicador_prov,est_indicador_depa),
est_gamlss=c(est_gamlss,est_gamlss_prov,est_gamlss_depa),
prom_Y_area=c(prom_Y_area,prom_Y_area_prov,prom_Y_area_depa))

ggplot(datos_est_nj, aes(x=est_indicador, y=prom_Y_area, color=nj_intervalos,shape=niv_geo)) +
geom_point(size=4,alpha = 7/10)+ geom_abline(slope=1, intercept=0, col = "red") +
xlab("Estimaciones directas") + ylab("Valores poblacionales") + ylim(1, 5) + xlim(1,5)+
labs(color = "Tamaños de muestra",shape="División geográfica") +
```

```
theme(legend.position = "none")

ggplot(datos_est_nj, aes(x=est_gamlss, y=prom_Y_area, color=nj_intervalos,shape=niv_geo)) +
geom_point(size=4,alpha = 7/10)+ geom_abline(slope=1, intercept=0, col = "red") +
xlab("Estimaciones GAMLSS") + ylab("Valores poblacionales")+ ylim(1, 5) + xlim(1,5)+
labs(color = "Tamaños de muestra",shape="División geográfica")

#Grafico IC - GAMLSS (departamento, provincias y distritos)
plot(c(est_gamlss_depa,est_gamlss_prov,est_gamlss),pch=19,main = NULL,
xlab = "",ylab = "Valor estimado",ylim = c(1,5),xaxt = "n")
arrows(1:(1+5+J), c(IC_inf_gamlss_depa,IC_inf_gamlss_prov,IC_inf_gamlss),
1:(1+5+J), c(IC_sup_gamlss_depa,IC_sup_gamlss_prov,IC_sup_gamlss),
length=0.05, angle=90, code=3)
points(c(prom_Y_area_depa,prom_Y_area_prov,prom_Y_area),
col=rgb(red = 1, green = 0, blue = 0,alpha=0.6),pch=19)
legend("bottomright", c("Est. GAMLSS","Valor real"),pch=19,
col=c("black","firebrick1"),bty="n",cex=0.8)
axis(1, at=1:(1+5+J), labels=c("11",
DatosCenso2017_prov$ubigeo,
DatosCenso2017$ubigeo),las=3,cex.axis=0.8)
```