**Universidad de la República**
**Facultad de Ciencias Sociales**
**Unidad Multidisciplinaria**
**Programa de Población**

# Luminosidad nocturna: variable sintomática o auxiliar dasimétrica para estimaciones de población en áreas pequeñas.

## Caso Uruguay entre 1996 y 2011

Autor: Richard Detomasi

Tutores: Leandro González (UNC-Argentina)

Virginia Fernández (UdelaR-Uruguay)

Tribunal: Mathias Bourel, Juan José Goyeneche, Yuri Resnichenko

13 de Marzo de 2019

Detomasi Araujo, Richard
Luminosidad nocturna: variable sintomática o auxiliar dasimétrica para estimaciones de población en áreas pequeñas.Caso Uruguay entre 1996 y 2011/ Richard Detomasi Araujo
Tesis Maestría en Demografía y Estudios de Población.- Montevideo: UR.FCS. Unidad Multidisciplinaria. Programa de Población, 2018
71 h, gráfs. cuadros. (Serie Tesis de Maestría en Demografía y Estudios de Población,13)

Incluye bibliografía.

1. Tesis. 2. Luminosidad nocturna. 3. Estimaciones poblacionales. 4. Estimaciones en áreas pequeñas. 5. Uruguay. I. Título

**Resumen**

Esta investigación plantea evaluar el uso de la luminosidad nocturna a partir de imágenes satelitales, como variable sintomática para estimar la población entre censos en Uruguay por pequeñas áreas (píxeles de 1 km$^2$) en el período 1996-2011; así mismo analizar esta variable como dasimétrica de la información de consumo de energía eléctrica residencial con el mismo fin. Se modeló su relación mediante funciones tales como regresiones lineales, GAM, CART, Bagging y Random Forest, comparándose en cada escenario en busca del estimador óptimo. Finalmente, se analizó la precisión de las estimaciones de ambos métodos, a partir de la comparación de los resultados con las estimaciones oficiales para el período a nivel nacional (Uruguay) y subnacional (en los 19 departamentos). Posteriormente, dichas estimaciones fueron comparadas con los datos a nivel de píxel de 1 km$^2$ del Censo1996 y el Conteo 2004.

**Palabras clave:** Luminosidad nocturna - Estimaciones poblacionales - Estimaciones en áreas pequeñas - Uruguay

**Abstract**

This research aims to test two methods based on nocturnal luminosity-based methods derived from satellite images to estimate the population dynamics between national and regional-wide censuses carried on small areas (pixels of 1 km$^2$) within Uruguay, in the period ranging from 1996 to 2011. On the one hand, luminosity is used as a symptomatic variable to perform the estimation. On the other hand, luminosity is used as a dasymetric variable for estimating the residential electrical energy consumption to perform the estimation. The relationship between these variables was explored and modeled by using linear regression, GAM, CART, Bagging and Random Forest functions, and each scenario was then compared in search of the optimal estimator. The precision of the estimates resulting from both methods was further analyzed, by comparing the optimal estimator resulting values with the officially published estimates for the period, at both the national and departmental (i.e. regional) levels. Finally, the obtained values were compared at the pixel-level scale (1 km$^2$ area) with available data from the 1996-Uruguayan Census and the 2004 Population Count.

**Keywords:** Night Luminosity - Population estimation - small area estimations - Uruguay

# 1. Agradecimientos

En primer lugar debo mi gratitud a mis tutores, Leandro González, por su esfuerzo y dedicación al guiarme a pesar de la distancia en las tierras de las estimaciones de población; y Virginia Fernández, por sus certeros comentarios y valiosos aportes.

También me gustaría reconocer, a Luiz Felipe Barros, otro aventurero de la luminosidad nocturna, y cuya tesis me iluminó cuando todo parecía oscuro.

Agradezco a Wanda Cabella e Igancio Pardo, por dedicarme su tiempo en pos de la culminación de esta etapa.

A Germán Botto, por la disposición permanente al intercambio, capacidad y generosidad, y a Nicolás Batalla, por amigo, hermano y apoyo incondicional.

Dejo para el final lo más importante, mi familia, sin cuyo cariño y apoyo no hubiera podido llegar hasta aquí. En particular a Ana, Eugenia y Matteo, que me supieron soportar a pesar de los duros vientos, sin dejar quebrar el mástil de esta nave familiar que nos abriga; mi madre y hermana, que desde la calma y la paciencia supieron acompañarme; mi padre, que desde mí vive, y en mí mantiene vigente su búsqueda del sano saber de la verdad.

# Índice

# Índice de figuras

# Índice de cuadros

# 2. Fundamentación

## 2.1. Introducción

Las estimaciones de población son una de las técnicas demográficas más importantes para la toma de decisiones. Con ellas, se tiene una referencia cuantitativa para distribuir los recursos del Estado de una forma eficiente. Por su parte, las estimaciones de población en áreas pequeñas, son un tradicional desafío metodológico para la demografía, dadas las dificultades que presentan para ello el uso de las herramientas más habituales con las que se modela la dinámica demográfica. El desarrollo tecnológico y de fuentes informacionales en los últimos tiempos, permiten hoy evaluar las diversas herramientas metodológicas en materia de estimación a pequeña escala de poblaciones en general, pero al momento de aplicación de las mismas, en función de las fuentes y las unidades de análisis propias de cada territorio, comienzan a configurarse particularidades que deben ser analizadas puntualmente.

Cabrera (2011) en su análisis de posibles variables sintomáticas para estimar la población intercensal en Uruguay, presenta al consumo residencial de energía eléctrica como la variable que mejor ajusta a las estimaciones hechas por el método de componentes por parte del Instituto Nacional de Estadística del Uruguay (INE). En esta investigación, basado en este antecedente, lo que se plantea es la posibilidad de utilizar imágenes satelitales de iluminación nocturna del Operational Linescan System (OLS) del Defense Meteorological Satellite Program (DMSP), denominadas Nighttime Lights Global Composites (Version 4), para los respectivos años intercensales considerándolas en un primer escenario como una variable sintomática de la cantidad de población, y en un segundo escenario utilizarlas como variable auxiliar dasimétrica para discretizar la información del consumo residencial de energía eléctrica a nivel nacional, de la Administración Nacional de Usinas y Transmisiones Eléctricas (UTE), que se encuentra a escala de las Oficinas Comerciales (OC).

El uso de imágenes satelitales fue considerado, dado que entre los productos globales de datos de teledetección, presentan la más alta correlación con las actividades humanas. Se han utilizado, en estudios sobre población (Doll, 2010), consumo de energía (Kiran Chad et al., 2009), PIB (Zhao et al., 2011; Sutton et al., 2007), movimientos estacionales de población asociado a actividad zafral (Bharti et al., 2011), cartografía de límites de ciudad (Imhoff et al., 2010; Sutton et al., 2010), volumen de quema de gas (Elvidge et al., 2009) y emisiones de CO2 (Oda et al., 2011; Feng-Chi et al., 2013), entre otros. Y son particularmente interesantes para esta

investigación, los trabajos que han establecido vínculos entre la extensión del área iluminada y el recuento de la población, la densidad de población y/o la distribución de la población (Sutton, 1997; Sutton et al., 1997; Pozzi et al., 2003; Dória, 2015).

Elvidge et al. (1997:14) propusieron un flujo de trabajo sistemático para fabricar composiciones de la luminosidad nocturna a partir de observaciones del OLS del DMSP, y desde 1994, el National Geophysical Data Center (NGDC) ha estado produciendo anualmente conjuntos de datos de luminosidad nocturna a nivel global. También se han desarrollado algoritmos automáticos para evaluar la calidad de las observaciones nocturnas de la banda visible para eliminar áreas con propiedades indeseables, como la contaminación por la luz solar o la luz de la luna, presencia de nubes, luces procedentes de fuentes efímeras y ruido de fondo (Elvidge et al., 1997). Usaremos los productos "avg_lights_x_pct" de 1996 a 2011, que actualmente se distribuyen en el sitio web del NGDC.

## 2.2. Problema de investigación

El problema de investigación de esta tesis, se enmarca en los estudios de estimación y proyección de población en áreas pequeñas, dado que se busca estimar la población para los años intercensales para todo el territorio uruguayo, a escala de píxeles de 1 $km^2$, analizando el potencial uso de imágenes satelitales de luminosidad nocturna per se, o como función de distribución de los datos de consumo eléctrico residencial.

Esta investigación profundizó en las dificultades expuestas por Cabrera (2011), quien evaluó varias posibles variables sintomáticas para las estimaciones intercensales de población en Uruguay, dando como la de mejor performance a la información de potencia residencial consumida de UTE. Para optimizar estos análisis que fueron realizados a escala de departamentos, es que en esta oportunidad se plantea evaluar, auxiliados por las imágenes satelitales de luminosidad nocturna, dos posibles aplicaciones de las mismas en materia de estimación para áreas pequeñas. En una primera instancia, se evalúa esta información como variable sintomática para la estimación intercensal de población en áreas pequeñas en el caso de Uruguay; para en una segunda instancia, evaluarla como variable auxiliar dasimétrica de la información de consumo eléctrico residencialen en el territorio.

En definitiva, el objetivo general de este proyecto fue el de desarrollar una aproximación a las estimaciones de población en áreas pequeñas, incorporando el uso de la luminosidad nocturna como variable sintomática y/o auxiliar dasimétrica del consumo eléctrico residencial. De acuerdo con este objetivo general, los objetivos

específicos fueron: (1) Sistematizar la información y conocimiento existentes en materia de metodologías de estimación de poblaciones en áreas pequeñas; (2) Desarrollar, describir y aplicar las metodologías consideradas al caso uruguayo, para las características específicas de las estimaciones de población; y (3) determinar los aportes que pueden extraerse de estas metodologías de cara al desarrollo de futuras estimaciones similares y de aquellas que quieran integrar la perspectiva de las áreas pequeñas, para el caso uruguayo en particular, y de forma general para la región.

En línea con estos objetivos, las preguntas de investigación que han orientado esta investigación, corresponden a cuestionarse cuán óptimo es el uso de la luminosidad nocturna como una variable sintomática relevante para la estimación de población en áreas pequeñas; al mismo tiempo que, cuál es el peso de la luminosidad nocturna para la discretización de la potencia consumida, en las estimaciones de poblaciones en áreas pequeñas.

# 3. Antecedentes y marco teórico

## 3.1. Introducción

En este capítulo se realiza un trayecto en etapas por los antecedentes que enmarcan esta investigación. A saber, los de las estimaciones poblacionales, especificándose lo relacionado a las estimaciones en áreas pequeñas; por otra parte, el uso de imágenes satelitales, y su devenir con mejoras constantes tanto en lo técnico como en su aplicación a diversas líneas de investigación. En una tercera instancia se desarrollan los conceptos específicos de variables sintomáticas y variables auxiliares dasimétricas; y finalmente, dadas estas dos opciones para el uso de la luminosidad nocturna, es que se presentan las bases de los modelados que se llevaron a cabo en estos dos escenarios.

## 3.2. Estimaciones de población

Desde las primeras proyecciones para Inglaterra y Gales realizadas por Edwin Cannan en 1885, el método de componentes o de cohorte-componentes ha sido la principal técnica utilizada para estimaciones poblacionales.

A nivel de metodologías para la estimación de poblaciones, hay que distinguir primero entre las sistematizaciones que de las mismas han surgido. Una clasificación inicial se realiza en relación al tiempo que refieren, pudiendo corresponder a estimaciones precensales para estudios de demografía histórica, intercensales o postcensales (Bryan, 2004 :523).

Otro esquema, más general, clasifica los métodos de estimación en dos tipos: (1) "flujo" y (2) "stock" (Long, 1993). En general, las acciones suelen tener un cierto stock en un punto en el tiempo (por ejemplo, el tamaño de la población en 2011 por el método de razón censal), mientras que un flujo (o "tasa") cambia una acción en el tiempo (i.e. la técnica de los componentes, que estima que cada componente del cambio de población desde el último censo); acciones y flujos son los bloques básicos de construcción de modelos de dinámica de sistemas.

Judson y Swanson (2011:13-14) plantean un esquema que categoriza las metodologías en modelos analíticos y estadísticos, matemáticos, y basados en muestras. González y Torres (2012) por ejemplo, las separan en: de componentes, matemáticas y sintomáticas; Murdock y Ellis (1991:181) por su parte lo hace con mayor grado de

desagregación, diferenciando entre aquellas de extrapolación y de razón, sintomáticas, basada en regresiones, y basadas en componentes, a las que habría que agregar una categoría "otras" que alberga a un variado grupo de otras metodologías (i.e. basados en muestras, registros administrativos o análisis por redes sociales).

Los antecedentes regionales compilados por González y Torres (2012), concentran ejemplos para el método de variables sintomáticas. Entre ellos el caso de Bay (1998) para Chile y Costa Rica, o el de Texeira Jardim (2001), quien aplica métodos de actualización de la población de los municipios de Río Grande do Sul (Brasil) durante la década de 1990. De modo similar, González y Torres (2012) mencionan el trabajo de Chávez Esquivel (2001) para los cantones de Costa Rica en los años 1990, en este caso aplicando los métodos de razón censal, diferencia de tasas, correlación de razón y de correlación de tasa. Por otra parte, González y Torres (2012) también presentan métodos integrados, como el de Jannuzzi (2005) que emplea proyecciones por componentes para un nivel regional y un sistema de ecuaciones diferenciales para áreas municipales, a partir de un modelo de especies competitivas provenientes de la Ecología.

Dos antecedentes más cercanos, serían las tesis de Dória (2015) y Barros (2017). Dado que Dória (2015), siguiendo trabajos anteriores que observaron relaciones lineales entre variables socioeconómicas y luminosidad nocturna (Elvidge et al., 2001; Amaral et al., 2005; Doll et al., 2006), utilizó el método de regresión lineal simple, con los datos de población como variable de respuesta, y las sumas totales de los números digitales (ND) de los píxeles de cada mancha de luz como variables explicativas, para la estimación poblacional a escala regional, en el Distrito Forestal Sustentable de la BR163, y en la escala local, estudiando la Región Metropolitana de San Pablo,en los años censales 2000 y 2010, además del conteo 2007. Por su parte, Barros (2017) analizó como variables sintomáticas de la población, varias fuentes de registros administrativos y la luminosidad nocturna, realizando un análisis transversal y longitudinal, ajustando modelos de regresión lineal múltiple, a nivel de municipio para todo Brasil.

Históricamente en Uruguay se han realizado las estimaciones y proyecciones poblacionales por la metodología de componentes (INE, 1998, 1999, 2005, 2014; DGEC, 1991; CELADE, 1981, 1984). Con el pasar del tiempo se ha ido ajustando, pasando de tramos quinquenales a edades simples, e incluso se han elaborado con niveles de desagregación, proyecciones subnacionales, utilizando categorías como "urbano" y "rural" (DGEC, 1991; INE, 1998); o Montevideo y "resto urbano" (INE, 1999). En estos casos se diferenció lo "urbano" de lo "rural" por departamento, utilizando

herramientas informáticas (Rural Urban Projection (RUP) y para datos agregados de regiones subnacionales su versión RUPAGG) brindadas por el US Census Bureau (INE, 2005), aunque para el último relevamiento no fueron utilizadas.

De todos modos, estando ya publicadas las proyecciones poblacionales a nivel nacional (INE, 2014), se encuentran en proceso de publicación los resultados desagregados aplicando el método de componentes, utilizando un enfoque de tipo multiregional, lo que permite diferenciar en cada departamento las categorías Urbano mayor de 5000 habitantes, urbano menor de 5000 habitantes y "rural" (Nathan com. pers., 2016).

A su vez, a nivel nacional, y fuera de las publicaciones en el marco del INE, se deben destacar el trabajo de Calvo y Prats (1992) sobre la proyección de la población de Canelones, utilizando la metodología de relación de cohortes (Duchesne, 1988), y principalmente, el trabajo de Cabrera (2011) que presenta un intento de construir proyecciones por un método basado en variables sintomáticas.

Cabrera (2011) menciona a su vez la existencia de una experiencia de elaboración de proyecciones a escala de localidades realizada por Calvo y Rios (1998) como insumo de un estudio en convenio de UTE con la Universidad de la República, que abarca el período 2000-2030. Para ese estudio se tomaron como base las proyecciones nacionales realizadas por el método de los componentes y se aplicó el método de Duchesne (1988), para estimar la población por departamentos; asimismo el método del parque habitacional[1] fue utilizado para proyectar secciones y localidades, manteniendo el supuesto de que no hay diferencias en el movimiento migratorio entre las distintas localidades del departamento.

Cabe aclarar que se trabajó con métodos de estimación y no de proyección de población, es decir, no se planteo realizar supuestos sobre el comportamiento futuro de cada componente de la dinámica (nacimientos, fallecimientos y migraciones) sino que se utilizó información simultánea al momento de la estimación para modelar la misma, por más que en algunos modelos se incorpora información pasada, bajo el supuesto que determinadas relaciones que se constataron durante el período anterior, se mantendrán estables hasta el momento de la estimación. (Cabrera, 2011 :19)

[1]El método de parque habitacional, considera las viviendas particulares existentes al último censo y la información sobre permisos de construcción de viviendas particulares.

## 3.3. Imágenes satelitales

Como ya se ha mencionado, las imágenes que se utilizaron corresponden al Operational Linescan System (OLS) del Defense Meteorological Satellite Program (DMSP), de las que se tomó la versión 4, de iluminación estable, con píxeles de 1 km$^2$ para todo Uruguay. En concreto se utilizó el producto de luminosidad nocturna "avg_lights_x_pct", que deriva del promedio del número digital (ND) de la banda visible de detecciones de luz libre de nubes, multiplicado por el porcentaje de frecuencia de detección de luz (Elvidge et al., 1997).

La inclusión de la frecuencia en términos de porcentaje de detección, normaliza los valores digitales resultantes para las variaciones en la persistencia de la iluminación, con un rango [0-63]. Por ejemplo, el valor de una luz que solo se detecta la mitad del tiempo es descontado en un 50 %. Debe tenerse en cuenta que este producto contiene detecciones de incendios y una cantidad variable de ruido de fondo.

También es relevante considerar que los relevamientos de estas imágenes satelitales fueron llevados a cabo por 5 satélites, que en períodos duplican los datos para ciertos años, por lo que se promediaron sus valores por píxel en el marco del procesamiento, obteniendo así una banda de luminosidad por año.[2]

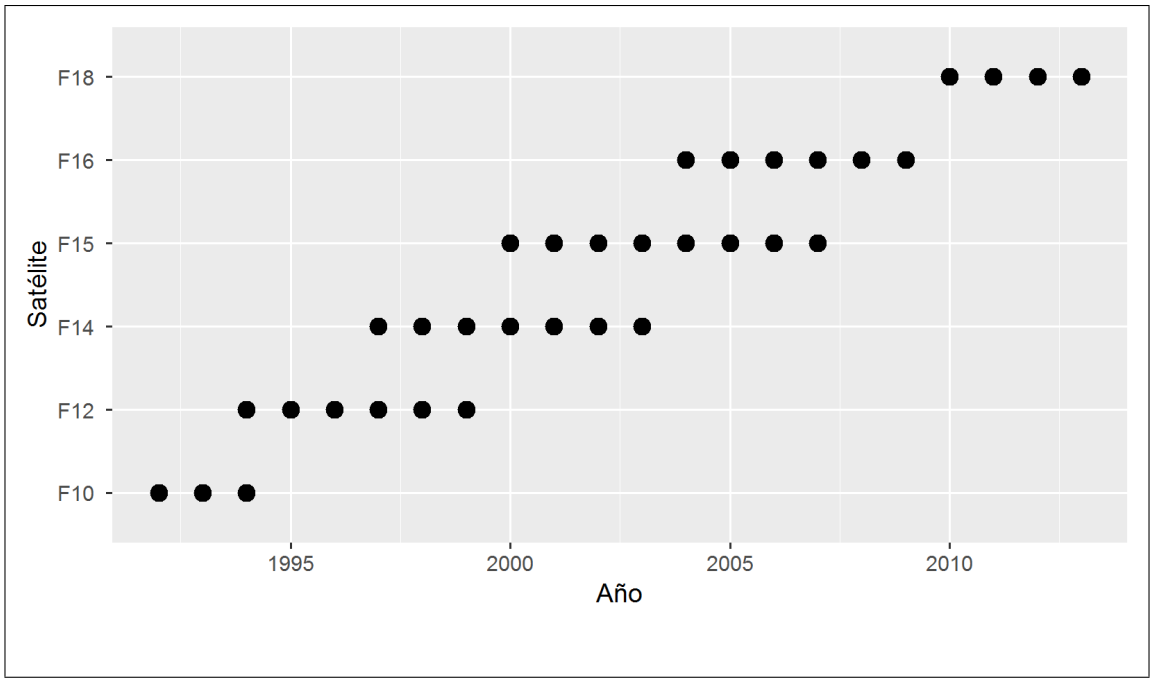


**Figura 3.1:** Serie histórica completa de los satélites DMSP con sensores OLS. Fuente:National Oceanic and Atmospheric Administration's. (NOAA).

[2]Disponible en: https://ngdc.noaa.gov/eog/dmsp/downloadV4composites.html.

Este proceso de sustitución de satélites, trae acarreado que los datos de luminosidad nocturna provenientes de la serie DMSP/OLS, precisen ser intercalibrados debido a la degradación de los sensores; sobre todo considerando la extensión temporal del período analizado.

Se han propuesto varios métodos para superar la falta de calibración entre satélites. Estos incluyen la región invariante y el método de regresión cuadrática propuesto por Elvidge et al. (2009), el método de regresión de segundo orden y umbral óptimo propuesto por Liu et al. (2012), y un método de regresión de ley de poder propuesto por Wu et al. (2013). Aunque los estudios basados en estos métodos de calibración mostraron una mejora del rendimiento después de la rectificación (Liu et al., 2012, Wu et al. (2013)), la suposición de que la luz nocturna permanece estable a lo largo del tiempo en un área en particular requiere una elección cuidadosa de la región invariante manualmente.

En esta investigación, para la debida adaptación al caso de Uruguay, se tomó el método definido por Elvidge et al. (2009), quienes realizan su intercalibración para Italia, pero solo para el período 1994 a 2008 (F121994 a F162008) y tomando como referencia para el ajuste de los modelos de regresión basado en los datos del satélite F12 de 1999, teniendo como área de referencia a Sicilia, Italia. Se ha elegido este método, ya que ha tenido una amplia aplicación, con eficacia comprobada (Elvidge y Sutton, 2011; Small y Elvidge, 2013; Han et al., 2014; Dória, 2015; Bennett y Smith, 2017)

Otra aplicación de esta misma metodología corresponde a Barros (2017), quien realizó la intercalibración para toda la serie desde 1992 a 2013 (F101992 a F182013), tomando como base el mosaico de referencia el del satélite F18, para el año 2012, usando como área de referencia a Salvador de Bahia, Brasil.

A nivel metodológico, el primer paso para la intercalibración según la metodología de Elvidge et al. (2009), es escoger un mosaico de referencia para el caso de Uruguay, que de acuerdo con los autores es aquel que presenta los mayores NDs y de mayor número de píxels saturados, (o sea, píxeles con ND=63). Para los datos de Uruguay, el satélite F18 en el año 2011 fue el que mejor atendió esos requisitos.

El segundo paso, corresponde a ajustar una función de segundo grado para realizar la intercalibración entre las imágenes. donde $ND_{ajustado}$ es el valor del píxel ajustado, $C_0$, $C_1$ y $C_2$ son los coeficientes y $ND$ es el valor original del píxel

$$ND_{ajustado} = C_0 + C_1 ND + C_2 ND^2$$

En tanto el tercer, y último paso, para la obtención de la intercalibración propuesta por Elvidge et al. (2009), es el cálculo de la media aritmética de los valores de los NDs para los años en que había más de un satélite captando la información.

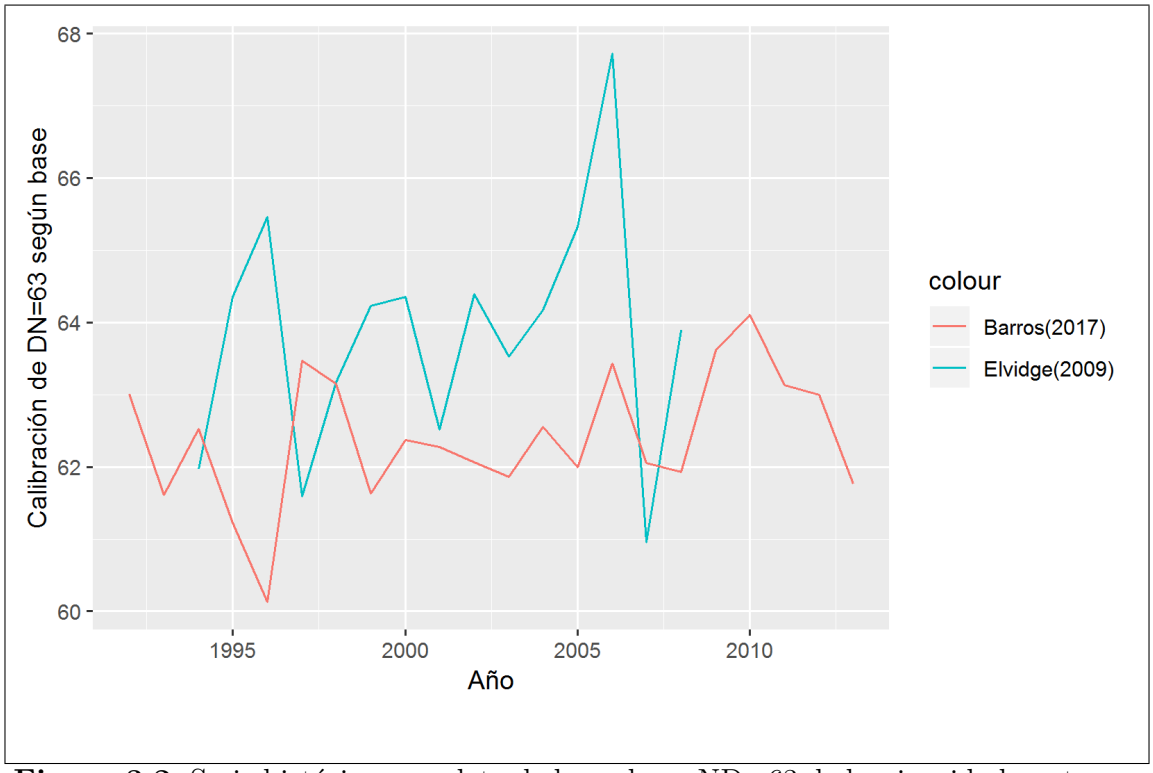


**Figura 3.2:** Serie histórica completa de los valores ND=63 de luminosidad nocturna, según los cálculos de intercalibración. Fuente: Basado en Elvidge et al. (2009) y Barros (2017).

Como concluye Barros (2017) el proceso de intercalibración, si bien es necesario para garantizar la comparabilidad de los datos, no garantiza una mejora en las relaciones entre la evolución de la luminosidad nocturna y de la población.

## 3.4. Cambiando la escala: modelando distribución de población

En los anteriores apartados, se ha constatado la utilización de áreas menores correspondientes a unidades de relevamiento de las instituciones encargadas de los Censos, pero en esta tesis se procura realizar una aproximación al territorio con el mayor nivel de detalle que permita la herramienta, para conocer mejor la distribución de la población.

Es por esto que se ha optado por trasformar los datos de población sobre un modelo

de distribución espacial de la misma, basado en la georreferenciación de hogares mediante información disponible en el SIIAS ajustada a 2011, integrando datos del INE, UTE y MIDES, es que se modeló la distribución de hogares. Luego se modeló la distribución de la población a 2011, sorteando proporcionalmente por segmento censal dentro de la capa anterior de puntos compuesta por hogares. Con estos insumos es que posteriormente se agregó este dato dentro de la grilla de 1 $km^2$ correspondiente a los píxeles de las imaágenes satelitales utilizadas.

Es importante resaltar que los ejercicios mostrados de desagregación espacial de la población posibilitan una mejor captación de las inequidades territoriales al comparar con las desagregaciones realizadas en otros niveles. Este potencial es altamente valioso si se pretende planificar políticas públicas desde una perspectiva espacial.

## 3.5. Variables sintomáticas y variables auxiliares dasimétricas

El punto tratado en este apartado, corresponde a las categorías de uso de la variable en la que se centra esta investigación, por tanto es imprescindible aclarar cuál es el significado que le daremos a cada término.

Un indicador o variable sintomática es una variable que cambia en el tiempo en concordancia con cambios en el volumen de la población en el mismo período; mientras que un indicador o variable auxiliar dasimétrica es una variable que cambia en el espacio en concordancia con cambios en la distribución de la población en el mismo período.

Más concretamente, en el marco de las estimaciones demográficas, hay que distinguir este análisis, entre los métodos matemáticos, donde las llamadas *variables sintomáticas*, siguiendo a Howe (2004:3) y CELADE (1998:78-79), se puede definir, como un conjunto o serie de datos que muestran alta correlación con los cambios en el tamaño de una población, vinculándose generalmente con registros estadísticos asociados al volumen y cambio de una población.

Mientras que para el segundo escenario, se debe considerar que el modelado dasimétrico que incluye un conjunto de técnicas para representar con mayor precisión la distribución espacial de una población dentro de las regiones espacialmente agregadas (Slocum y Slocum, 2009, capítulo 15). Los datos espaciales auxiliares son esenciales para el proceso dasimétrico. Los modeladores de dasimetrías a menudo clasifican los

datos auxiliares en dos tipos: limitaciones y variables auxiliares relacionadas.

Las variables limitantes establecen restricciones en los valores de población permisibles, por ejemplo, limitando la cantidad de población a cero en áreas cubiertas por agua. Las variables auxiliares relacionadas pueden acomodar relaciones más complejas. Por ejemplo, la densidad de la carretera, la elevación o la cobertura de la tierra pueden usarse para amplificar o limitar las densidades de la población (Mennis, 2009).

## 3.6. Modelados

### 3.6.1. Introducción

Desde que a comienzos del siglo XIX, Legendre y Gauss publicaron artículos sobre el método de los mínimos cuadrados, se implementaron de la forma más temprana lo que ahora se conoce como regresión lineal (Legendre, 1805; Gauss, 1809, 1821). Su uso se vincula a predecir valores cuantitativos, como el salario de un individuo. Para predecir los valores cualitativos, como si un paciente sobrevive o muere, o si el mercado de valores aumenta o disminuye, Fisher (1936) por su parte, propuso el análisis discriminante lineal. En la década de 1940, varios autores presentaron un enfoque alternativo, la regresión logística. A principios de la década de 1970, Nelder y Wedderburn (1972) acuñaron el término modelos lineales generalizados para una clase completa de métodos de aprendizaje estadístico que incluyen regresión lineal y logística como casos especiales.

Hacia el final de la década de 1970, muchas más técnicas para aprender de los datos estaban disponibles. Sin embargo, eran casi exclusivamente métodos lineales, porque la adaptación de relaciones no lineales era computacionalmente inviable en ese momento. En la década de 1980, la tecnología informática finalmente había mejorado lo suficiente como para que los métodos no lineales ya no fueran computacionalmente prohibitivos. A mediados de la década de 1980, Breiman et al. (1984) introdujeron los árboles de clasificación y regresión, y fueron de los primeros en demostrar el poder de una implementación práctica detallada de un método, incluida la validación cruzada para la selección del modelo. Hastie y Tibshirani (1986) acuñaron el término modelos aditivos generalizados (GAM por sus siglas en ingles), para una clase de extensiones no lineales a modelos lineales generalizados, y también proporcionaron una implementación de software práctica.

Desde entonces, inspirado por el advenimiento del aprendizaje automático y otras

disciplinas, el aprendizaje estadístico ha surgido como un nuevo subcampo en la estadística, centrado en el modelado y la predicción supervisados y no supervisados[3]. En los últimos años, el progreso en el aprendizaje estadístico se ha caracterizado por la disponibilidad cada vez mayor de software potente y relativamente fácil de usar, como el sistema R popular y de libre disponibilidad. Esto tiene el potencial de continuar la transformación del campo de un conjunto de técnicas utilizadas y desarrolladas por estadísticos e informáticos a un conjunto de herramientas esenciales para una comunidad mucho más amplia.

Según Breiman (2001) hay dos culturas en el uso del modelado estadístico para llegar a conclusiones a partir de los datos. Uno que supone que los datos son generados por un modelo de datos estocástico dado y otro que usa modelos algorítmicos y trata el mecanismo de datos como desconocido. Gráficamente se representa por los siguientes tres esquemas.

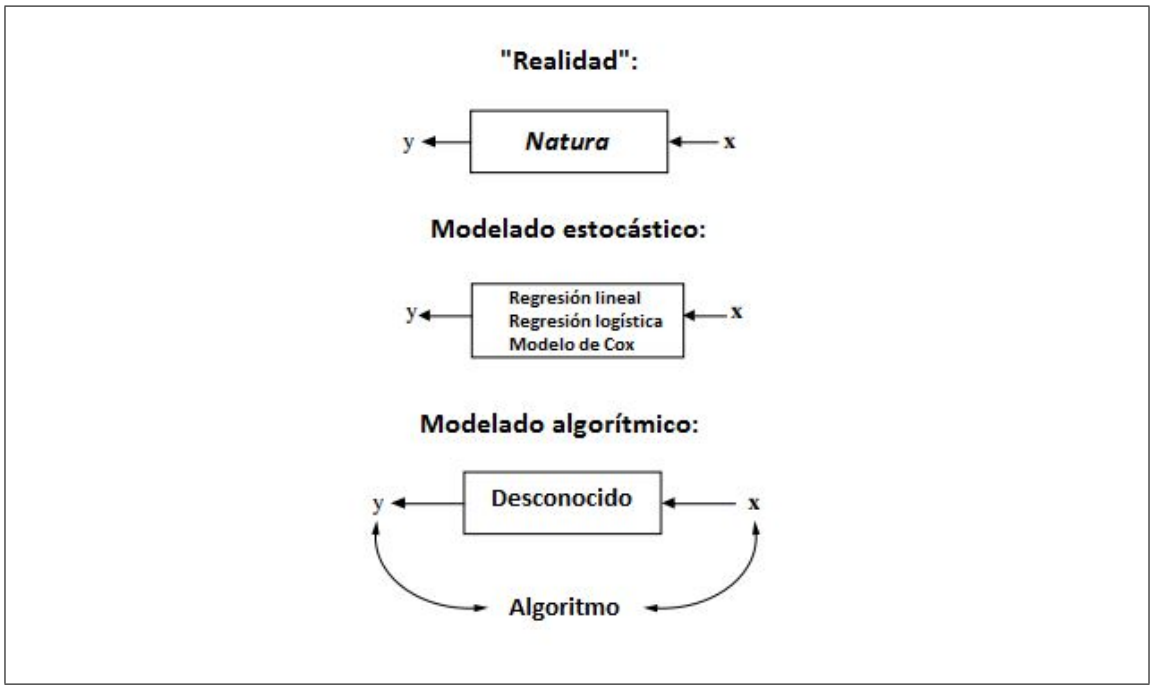


**Figura 3.3:** Diagramas de culturas de modelado. Fuente: Breiman (2001).

Con la mira en aproximarnos a la "realidad", iniciaremos la presentación de los modelos estocásticos paramétricos que, traen consigo la posibilidad de que la forma funcional utilizada para estimar $f(x)$ sea muy diferente de la $f(x)$ verdadera, en cuyo caso el modelo resultante no encajará bien con los datos. Por el contrario, los

[3]El aprendizaje no supervisado es un método de Aprendizaje Automático donde un modelo es ajustado a las observaciones. Se distingue del Aprendizaje Supervisado por el hecho de que no hay un conocimiento a priori.

enfoques no paramétricos evitan por completo este peligro, ya que esencialmente no se hace ninguna suposición sobre la forma de $f(x)$.

Los métodos no paramétricos no hacen suposiciones explícitas sobre la forma funcional de $f(x)$. En su lugar, buscan una estimación de $f(x)$ que se acerque lo más posible a los puntos de datos sin ser demasiado brusca. Dichos enfoques pueden tener una gran ventaja sobre los enfoques paramétricos: al evitar la suposición de una forma funcional particular para $f(x)$, tienen el potencial de ajustarse con precisión a un rango más amplio de formas posibles para $f(x)$. Una gran desventaja de esto es que al no reducir el problema de estimar $f(x)$ a un pequeño número de parámetros, se requiere una gran cantidad de observaciones, para obtener una estimación precisa de $f(x)$.

En esencia, el aprendizaje estadístico se refiere a un conjunto de enfoques para estimar $f(x)$, y aquí se esbozarán algunos de los conceptos teóricos clave que surgen en este proceso, así como las herramientas para evaluar las estimaciones obtenidas. La atención se centra en las técnicas para hallar $f(x)$ con el objetivo de minimizar el error reducible. Es importante tener en cuenta que el error irreductible siempre proporcionará un límite superior en la precisión de nuestra predicción para Y. Este límite casi siempre se desconoce en la práctica.

Los modelos algorítmicos por su parte, correspondientes a técnicas no paramétricas, logran una mejor adaptación a los datos disponibles, mediante la obtención de estimaciones más próximas a la curva de regresión subyacente, sin formular rígidos modelos paramétricos.

Otra forma de interpretar las diferencias entre los muchos modelos que se examinan en esta tesis, es que algunos son menos flexibles o más restrictivos, en el sentido de que pueden producir solo un rango relativamente pequeño de formas para estimar $f(x)$.

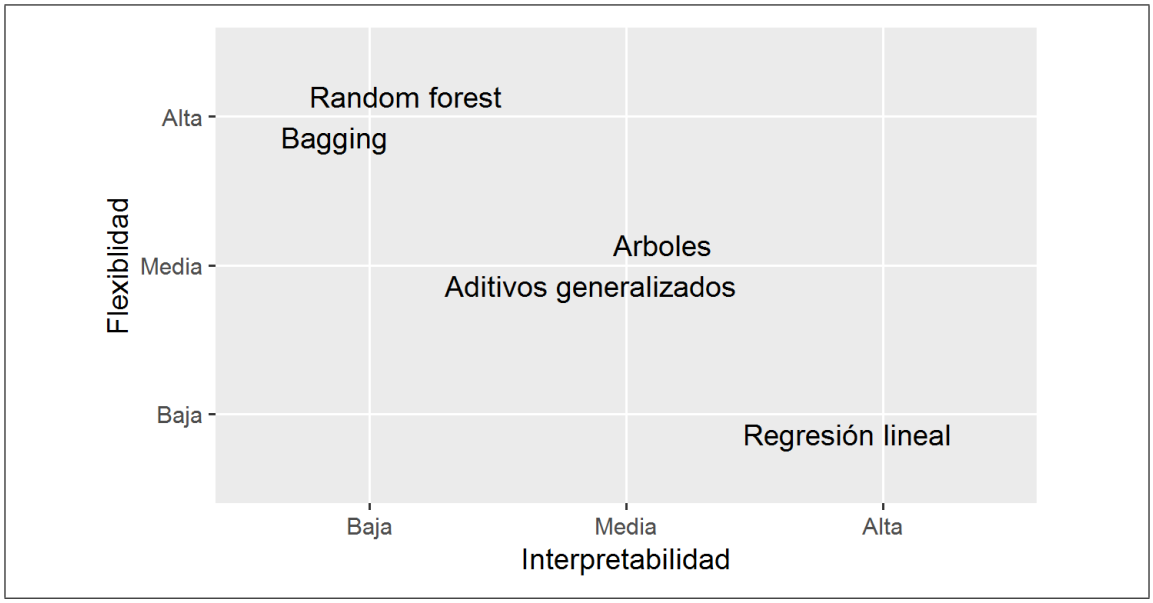


**Figura 3.4:** Equilibrio entre el poder explicativo (Interpretabilidad) y el poder predictivo (Flexibilidad) de los modelos. Fuente: Fuente: Basado en James et al. (2013).

Por ejemplo, la regresión lineal es un enfoque relativamente inflexible, ya que solo puede generar funciones lineales; otros métodos, como los modelos aditivos generalizados (GAM), son considerablemente más flexibles porque pueden generar un rango mucho más amplio de formas posibles para estimar $f(x)$; incluso en el caso de Random Forest, donde la flexibilidad es mayor aun, la interpretabilidad es sacrificada en el camino.

¿Pero por qué se elegiría usar un método más restrictivo en lugar de un enfoque muy flexible? Hay varias posibles razones para ello, principalmente si estamos interesados en la inferencia, dado que los modelos restrictivos son más interpretables. Por ejemplo, cuando la meta es la inferencia, el modelo lineal puede ser una buena opción, ya que será bastante fácil entender la relación entre $y$ y $x_1, x_2, ..., x_p$. Por el contrario, los enfoques muy flexibles, como las splines y los métodos de impulso, pueden llevar a estimaciones tan complicadas de $f(x)$ que es difícil comprender cómo se asocia cada predictor individual con la respuesta.

### 3.6.2. Modelos Lineales (LM)

Este tipo de modelo, ejemplo de los modelos paramétricos, reduce el problema de estimar $f(x)$ hasta estimar un conjunto de parámetros. Asumir una forma paramétrica para $f(x)$ simplifica el problema de estimar $f(x)$ porque en general

es mucho más fácil estimar un conjunto de parámetros, como $\beta_0, \beta_1, ..., \beta_p$ en el modelo lineal (2.4), que ajustar una función completamente arbitraria $f(x)$. La desventaja potencial de un enfoque paramétrico es que el modelo que elijamos generalmente no coincidirá con la verdadera forma desconocida de $f(x)$. Si el modelo elegido está muy lejos de la verdadera $f(x)$, entonces nuestro estimado será pobre. Podemos tratar de resolver este problema eligiendo modelos flexibles que puedan adaptarse a muchas formas funcionales posibles diferentes y flexibles para $f(x)$. Pero en general, ajustar un modelo más flexible requiere estimar un mayor número de parámetros. Estos modelos más complejos pueden conducir a un fenómeno conocido como sobreajuste de los datos, que esencialmente significa que el sobreajuste sigue las observaciones muy de cerca, lo que le quita capacidad para generalizar y adaptarse a una observación nueva.

El modelo lineal tiene claras ventajas en términos de inferencia y, en los problemas del mundo real, a menudo es sorprendentemente competitivo en relación con los métodos no lineales (James et al., 2013 :203)

### 3.6.3. Modelos Aditivos Generalizados (GAM)

En el marco de los modelos GAM, se tomaron dos casos dentro de la familia gaussiana, que toman como función link a la de identidad, y se consideraron dos penalizaciones, una basada en una regresión con *thin plate splines* (tp)[4], mientras que el Modelo 5 se basa en una penalización por *Cubic regression splines* (cr)[5]

### 3.6.4. Árboles de Regresión (CART)

La esencia de un árbol de regresión es que las funciones se particionan, comenzando con la primera división que mejora más la suma de cuadrados residual (SCR) . Estas divisiones binarias continúan hasta la terminación del árbol. Cada división

[4]Esta regresión se construye comenzando con la base y la penalización para una *thin plate splines* completa y luego truncando esta base de una manera óptima, para obtener un suavizador de rango bajo. Detalles en Wood (2003). Una ventaja clave del enfoque es que evita los problemas de colocación de nudos del modelado spline de regresión convencional, pero también tiene la ventaja de que los suavizados de menor rango se anidan dentro de rangos de mayor rango, por lo que es legítimo usar métodos convencionales de prueba de hipótesis. para comparar modelos basados en splines de regresión pura. Tenga en cuenta que el truncamiento de la base no cambia el significado de la penalización de *thin plate splines* (penaliza exactamente lo que habría penalizado por una *thin plate splines* completa).

[5]Por su parte una penalización por *Cubic regression splines* implica que se le ha modificado su penalización para reducirla a cero con parámetros de suavizado suficientemente altos. Se utilizan las bases *Cardinal spline*, véanse detalles completos Wood (2017 sección 5.3.1).

/ partición subsiguiente no se realiza en todo el conjunto de datos, sino solo en la parte de la división anterior que corresponde. Este proceso descendente se conoce como partición recursiva.

También es un proceso codicioso, un término con el que puede tropezar al leer acerca de los métodos de aprendizaje automático. Codicioso significa que, durante cada división en el proceso, el algoritmo busca la mayor reducción en la SCR sin tener en cuenta qué tan bien funcionará en las últimas particiones. El resultado es que puede terminar con un árbol completo de ramas innecesarias que conduce a un sesgo bajo pero una gran variación. Para controlar este efecto, debe podar apropiadamente el árbol a un tamaño óptimo después de construir un árbol completo (Lesmeister, 2017).

### 3.6.5. Métodos de Agregación de Modelos

Para mejorar en gran medida la capacidad predictiva de nuestro modelo, podemos producir numerosos árboles y combinar los resultados. Para esto surgen dos diferentes herramientas en el desarrollo del modelo por un lado la agregación por bootstrap o bagging que explicaremos a continuación. A la que la técnica del bosque aleatorio (Random Forest) Breiman (2001) agrega la selección aleatoria de atributos, introducida independientemente por Tin Kam Ho (1998).

#### 3.6.5.1. Bagging

En el bagging, un árbol individual se construye sobre una muestra aleatoria del conjunto de datos, aproximadamente dos tercios de las observaciones totales (nótese que el tercio restante se denomina out-of-bag (oob)). Esto se repite docenas o cientos de veces y los resultados se promedian. Cada uno de estos árboles crece y no se poda en función de una medida de error, lo que significa que la varianza de cada uno de estos árboles individuales es alta. Sin embargo, al promediar los resultados, puede reducir la varianza sin aumentar el sesgo.

#### 3.6.5.2. Random Forest (RF)

El random forest por su parte agrega que, al mismo tiempo que usa muestras aleatorias de los datos, es decir, bagging, también se toma una muestra aleatoria de las características de entrada en cada división. En el paquete randomForest(Liaw y Wiener, 2002b), fue utilizado el número aleatorio predeterminado de los predictores

que se muestrean, que, para la regresión como en este caso, es el número total de predictores dividido entre tres. El número de predictores que el algoritmo elige al azar en cada división se puede cambiar a través del proceso de ajuste del modelo (Lesmeister, 2017).

Los modelos de Random Forest (Breiman, 2001) son un enfoque conjunto de modelado no paramétrico que hace crecer un "bosque" de clasificación individual o árboles de regresión y mejora el Bagging Breiman (1996) utilizando lo mejor de una selección aleatoria de predictores en cada nodo de cada árbol (Breiman, 2001).

Este tipo de modelo consiste en muestrear $n$ árboles mediante bootstrap de los datos originales. Luego para cada muestra bootstrap, genera un árbol de regresión no podado, y en cada nodo, en lugar de elegir la mejor división entre todos los predictores, muestrea $m$ veces aleatoriamente la mayoría de los predictores y elije la mejor división entre esas variables (con el Bagging Se puede pensar en el caso especial de random forest obtenido cuando $m = p$, siendo $p$ el número de predictores). Finalmente, predice nuevos datos agregando las predicciones de los $n$ árboles promediando las estimaciones (Liaw y Wiener, 2002a).

Estos modelos tienen la ventaja de tener menos parámetros de ajuste. Y en nuestra metodología, esto es especialmente importante ya que la optimización se puede automatizar como parte del proceso de adaptación. Además, en el caso donde hay muchos predictores correlacionados o predictores con un amplio espectro de valor informativo (por ejemplo, algunos con mucha información entre muchos con muy poco), el algoritmo Random Forest es atractivo porque la salida del algoritmo de crecimiento forestal puede ser utilizado para estimar medidas de importancia de variables post hoc.

# 4. Marco metodológico

## 4.1. Introducción

En este capítulo, se detallará los recursos que fueron necesarios para la elaboración de esta tesis, además de especificar los procedimientos implicados en la toma de decisiones metodológicas que han hecho posible la combinación de las diversas fuentes, y compaginarlas a una misma escala de análisis, la de píxeles de 1 $km^2$.

## 4.2. Recursos necesarios

Los recursos utilizados, requirieron horas de procesamiento, y el relevamiento bibliográfico se basó en el amplio espectro de recursos que se disponibilizan por el Portal Timbó. En cuanto a las imágenes satelitales utilizadas están disponibles de forma abierta por el National Centers for Environmental Information (NCEI) del National Oceanic and Atmospheric Administration (NOAA). Finalmente, para el procesamiento fue utilizado software libre y abierto para todos los procesos, específicamente se usó lenguaje R (R Core Team, 2016) usando como interfaz Rstudio (RStudio Team, 2012), focalizándonos en varias de sus librerías complementarias, de igual modo liberadas (i.e. *sp* (Pebesma, 2005; Roger S. Bivand, 2013), y *raster* (Hijmans, 2016)).

Por otra parte, fue obtenido al amparo de la Ley 18381 de Acceso a la Información Pública, a través de la oficina de Registro e Información Documental de la Secretaría General de UTE: (1) los consumos residenciales de energía eléctrica para todo el país por Oficina comercial de UTE desde el año 1996 al 2017; y (2) los límites geográficos para todo el país de las áreas correspondientes a dichas Oficinas comerciales de UTE, corrigiéndose sus variaciones[6] entre 1996 y 2011. Datos anonimizados como corresponde según Ley 18.331 de Protección de Datos Personales.

Se contó además con bases de direcciones georreferenciadas para todo el país a 2011, compiladas por el Departamento de Geografía de la Dirección Nacional de Evaluación y Monitoreo (DINEM) del Ministerio de Desarrollo Social (MIDES), quienes complementando fuentes como los puestos de acceso a los servicios (luz y agua), las coordenadas de los relevamientos de la propia DINEM-MIDES y las

[6]Las oficinas tienen coberturas muy heterogéneas, estando conformadas por parte de localidades, más de una localidad o incluso cubren en algún caso áreas de más de un departamento. Para subsanar esta dificultad se gestionó con UTE los límites de las áreas cubiertas por cada oficina comercial.

coordenadas de los relevamientos rurales del Censo 2011 para los hogares rurales. Las mismas se utilizaron en ambos escenarios como variable dasimétrica limitante, dado que solo se estima en los píxeles donde a 2011 hay direcciones (i.e. plantas industriales en el medio rural, aeropuertos), como para sortear la cantidad de personas por segmento INE en el Censo 2011, entre las direcciones correspondientes al mismo segmento.

Esta última estrategia, corresponde a mejorar el modelado de la distribución poblacional a 2011 a escala del píxel de la imagen satelital 1km$^2$, aproximando la densidad de la población en segmentos INE de mayor área y que presentan diversos usos en su territorio (i.e. urbano consolidado, rural disperso, etc.) con sus propias variaciones de densidad.

## 4.3. Metodología

A partir de estos insumos, es que se compaginó a escala de píxel de 1km$^2$ para 2011, la luminosidad nocturna de los píxeles de la imagen satelital correspondiente a ese año transformadas a su logaritmo natural. De modo similar, con la información de consumo residencial discretizado a escala de 1km$^2$, se calculó el producto entre el consumo y logaritmo natural de la luminosidad nocturna para cada píxel.

Las cantidades poblacionales del Censo 2011, se tomaron de los datos del INE (2014), luego se sorteó aleatoriamente y con reposición por segmento censal INE, el $N$ de población censada entre los pares de coordenadas de direcciones gestionadas. Contabilizando por píxel, el $N$ de población sorteadas en cada uno de ellos. Este pasaje de la información a nivel de segmentos (polígonos), a pares de coordenadas (puntos), y luego agregados por píxel (poligono), contribuyó para tener un modelado de la distribución de la población a 2011 en la misma unidad que los datos de luminosidad nocturna, y por tanto fue utilizado para orientar la variable de respuesta para los modelos generados en ambos escenarios.

Se debe mencionar en este apartado, que se optó por desagregar el dato de población a la escala de los píxeles de las imágenes dado que cambiaron segmentos censales entre 1996-2004 pero no hubo cambios en sus códigos (resegmentaciones), cosa que sí ocurrió para el relevamiento 2011, y por ende trabajar a nivel de segmentos implicaría des-resegmentar las cantidades de este último relevamiento manejándonos con los segmentos censales usados hasta ese momento. De todos modos, se asume que diversas áreas de crecimiento urbano quedaron dentro de segmentos rurales en las comparaciones de las estimaciones por píxel para 1996-2004 frente a los conteos

respectivos, ya que fue a partir de estos que se evaluaron los resultados a esta escala.

Por otra parte, se elaboró para cada año entre 1996 y 2011, una densidad de consumo eléctrico residencial en una grilla regular con escala de $1km^2$. Distribuyendo dentro de los límites de cada oficina comercial de UTE la potencia residencial consumida, según los valores por píxel de luminosidad nocturna de las imágenes del año que correspondía.

Para cada escenario, se evaluaron: Modelos Lineales (LM), Modelos Aditivos Generalizados (GAM), Árboles de Regresión (CART), y Métodos de Agregación de Modelos como Bagging y Random Forest (RF) para cada una de estas variables sintomáticas frente a los datos de cantidad poblacional de 2011, comparando su calidad por validación cruzada. La comparación de performances de los modelos fue según criterio de información de Akaike (AIC) entre los modelos LM y GAM, y posteriormente se evaluó los errores y varianza explicada, sobre iteraciones de muestra de aprendizaje y de test de los 5 mejores modelos.

Finalmente, para evaluar el ajuste del modelo que optimice su performance tanto en muestras de aprendizaje como en las de test de cada escenario, se compararon dichas estimaciones para el período 1996-2010 agregadas a nivel de departamentos evaluándose sus errores en relación a las estimaciones oficiales hechas por métodos tradicionales. De modo similar, se evaluaron los errores de las estimaciones de dichos modelos, agregados los datos de población del Censo de 1996 y el conteo de 2004 a escala de píxeles.

# 5. Escenarios

## 5.1. Introducción

En este capítulo se presenta la evaluación de modelos; en primera instancia utilizando la información de luminosidad nocturna por píxel per se como Variable sintomática de población, para luego utilizarla como Variable auxiliar dasimétrica del consumo residencial de energía eléctrica, y a partir de esta combinación estimar la población.

Cabe mencionarse que en el primer escenario solo se utilizaron para todos los modelos, aquellos píxeles que cumplían dos condiciones: a) luminosidad mayor a 0 y b) población por píxel mayor a 0. Estas condiciones asumen que los valores 0 no se estiman, y que donde no había población en 2011 no habría población en los años anteriores.

En ambos escenarios, se aplicó logaritmo natural a los valores de luminosidad nocturna y de consumo residencial dasimetrizado por luminosidad nocturna, para expandir sus rangos dada la amplitud del rango de población por píxel en relación a la de ambas variables explicativas.

## 5.2. Modelos

Dado que se busca un modelo que aporte a la estimación de la cantidad poblacional ($y$), en el escenario Variable Sintomática (VS) la única variable utilizada fue el logaritmo del promedio de luminosidad nocturna ($x$), mientras que el escenario que llamaremos Variable Dasimétrica (VD) la variable utilizada para estimar la población fue el logaritmo natural del consumo residencial dasimetrizado por luminosidad nocturna.

En ambos escenarios, la primera clase de modelos que se evaluó fueron los lineales de primer, segundo y tercer grado; en concreto:

$$f_1(x) = \beta_0 x$$

$$f_2(x) = \beta_0 x + \beta_1 (x^2)$$

$$f_3(x) = \beta_0 x + \beta_1 (x^2) + \beta_2 (x^3)$$

Por su parte en el marco de los modelos GAM, se tomaron dos casos dentro de la familia gaussiana, que toman como función link a la de identidad, pero considerando en el Modelo 4 se basa en una regresión con *thin plate splines* (tp), mientras que el Modelo 5 se basa en una penalización por *Cubic regression splines* (cr), como se definen a continuación:

$$f_4(x) = \beta_0\ j_1(x) + \epsilon \text{ (basado en tp, con función link gaussiana)}$$

$$f_5(x) = \beta_0\ j_1(x) + \epsilon \text{ (basado en cr, con función link gaussiana)}$$

Como el objetivo es evaluar el mejor modelo de cada tipo de modelo, se compararon a la interna de cada uno de estos dos grupos cual presenta una mejor performance según el criterio de información de Akaike (AIC), para posteriormente agregarlos en la comparación por errores y varianza explicada, sobre iteraciones de muestra de aprendizaje y de test, a los mejores modelos de los próximos 3 tipos.

La tercer clase de modelos evaluados correspondió a Árboles de Regresión (CART), en este caso, probamos tres modelos, mediante la variación del parámetro de complejidad "$\alpha$"[7] de la función entre 0.01, 0.001 y 0.0001; podando en cada caso el árbol al más simple cuya bondad sea comparable al mínimo estimador con menor incertidumbre (regla 1-SE). Los modelos tuvieron la siguiente definición:

$$f_6(x) = \beta_0\ j_1(x) + \epsilon \text{ (con un } \alpha = 0{,}01)$$

$$f_7(x) = \beta_0\ j_1(x) + \epsilon \text{ (con un } \alpha = 0{,}001)$$

$$f_8(x) = \beta_0\ j_1(x) + \epsilon \text{ (con un } \alpha = 0{,}0001)$$

Esta progresividad entre los árboles propuestos corresponde a tener un primer árbol que aunque con pocas ramas, cumple perfectamente la función de descripción sin dificultad, en cambio, los que presentan mayor complejidad ($\alpha = 0.001$ o $0.0001$) pierden en descripción pero ganan en ajuste. La selección dentro de esta clase de modelos, fue mediante el cálculo del porcentaje de varianza explicada por cada modelo, dado que es una cantidad claramente interpretable. En definitiva, los tres

[7]El parámetro de complejidad ($\alpha$) implica que cualquier división que no disminuya la falta general de ajuste en un factor de $\alpha$ no se intenta. La principal función de este parámetro es ahorrar tiempo de cálculo mediante la poda de divisiones (Para profundizar véase Zhang y Singer (2010) y Therneau et al. (1997))

modelos podando según la regla 1-SE quedaron establecidos los siguientes valores de complejidad:

Cuadro 5.1: Valores del parámetro de complejidad para modelos según la regla 1-SE

| Modelo | $\alpha$ | $\alpha$ (1-SE) |
|---|---|---|
| 6 | 0.01 | 0.01 |
| 7 | 0.001 | 0.0014612 |
| 8 | 0.0001 | 0.00014465 |

El cuarto tipo de modelo evaluados, corresponde a Bagging, que considera para la construcción del estimador remuestras bootstrap de la base. Y dado que se basa en un CART definido, y nuestro objetivo es evaluar el mejor modelo para predecir, tomamos el CART con mayor complejidad de los anteriores ($\alpha = 0.00014465$). Evaluando en este tipo de modelo, los resultados con 25 o 100 remuestras, tomando el modelo que presente menor error de las observaciones *out of bag*.[8] La definición de estos dos modelos correspondió a:

$$f_9(x) = \beta_0\ j_1(x) + \epsilon \text{ (Con 25 baggs y un } \alpha = 0.00014465)$$

$$f_{10}(x) = \beta_0\ j_1(x) + \epsilon \text{ (Con 100 baggs y un } \alpha = 0.00014465)$$

Finalmente, el quinto tipo modelo trabajado, corresponde a Random Forest, dada la potencialidad de este tipo de modelo en relación a la actual capacidad de computo, se construyó uno con 200 árboles y otro con 500. Seleccionando el modelo que presentó mayor porcentaje de varianza explicada. Siendo la definición de estos modelos:

$$f_{11}(x) = \beta_0\ j_1(x) + \epsilon \text{ (Con 200 árboles)}$$

$$f_{12}(x) = \beta_0\ j_1(x) + \epsilon \text{ (Con 500 árboles)}$$

Cabe aclarar, que el uso de AIC, OOB y porcentaje de varianza explicada, para

[8]Out of bag (OOB) es un método para medir el error de predicción de modelos de aprendizaje automático ( *Bagging*, *Random Forest*, etc.), que utilizan la agregación de *bootstrap* para muestras de datos de submuestras utilizadas para el entrenamiento. OOB es el error de predicción promedio en cada muestra de entrenamiento $j_i$, utilizando solo los árboles que no tenían $j_i$ en su muestra de arranque.(James et al., 2013)

la evaluación diferencial de entre los modelos de cada grupo, viene dado por la asociación de cada grupo a una cierta herramienta de evaluación entre modelos. Pero cuando se evaluó los mejores modelos de cada escenario, se consideró la varianza explicada promedio por cada uno.

## 5.3. Resultados Variable Sintomática (VS)

El primer grupo correspondiente a modelos lineales, traen consigo el primer resultado, en tanto presentan una relación lineal entre las variables de luminosidad y cantidad poblacional evaluadas. Como muestra la Figura 5.1, entre los modelos LM, a mayor grado el ajuste mejora.

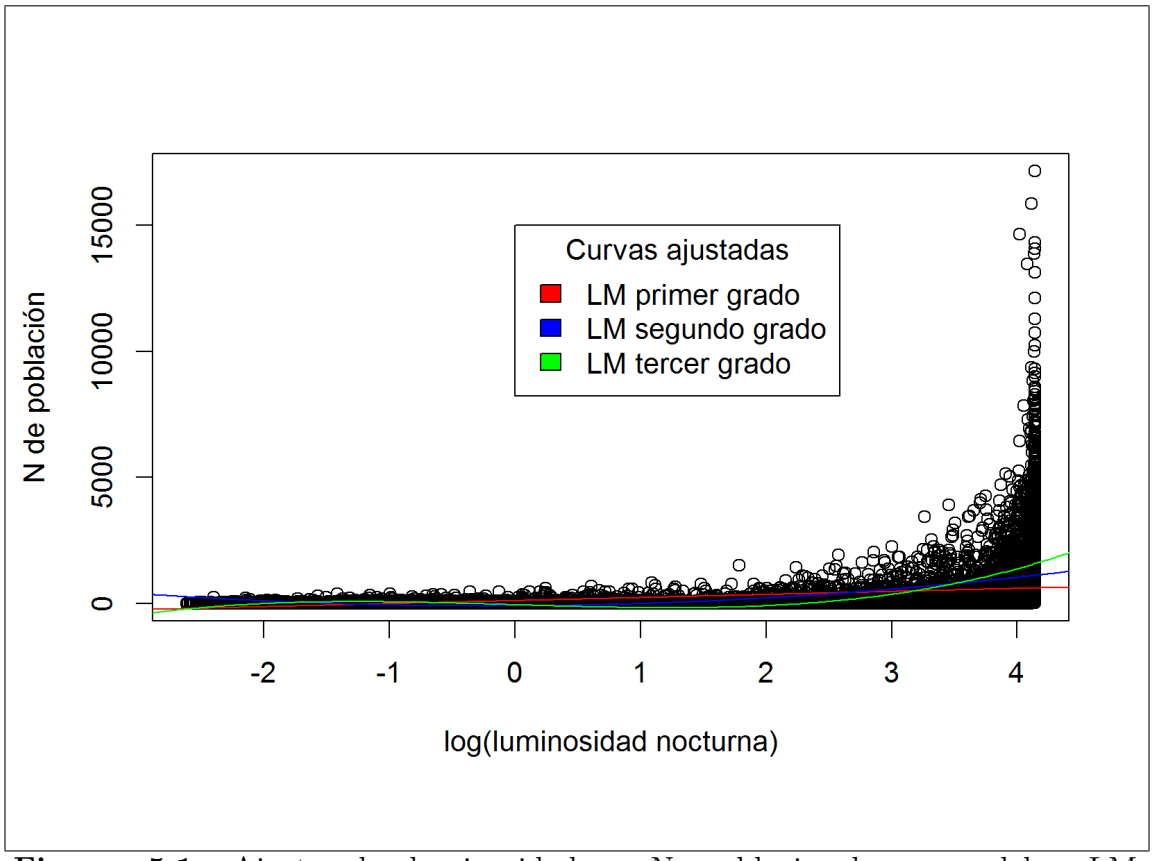


**Figura 5.1:** Ajuste de luminosidad y N poblacional en modelos LM. Fuente: Elaboración propia en base a procesamiento de imágenes satelitales OLS y marco censal 2011 (INE, 2014)

En cuanto a sus resultados en la comparación por AIC, el modelo 1 arrojó un AIC de 495954.42, mientras que el modelo 2 (segundo grado) obtuvo un AIC de 491807.16, finalmente el modelo 3 configuró un AIC de 487987.19; por tanto, el modelo LM que

se utilizó fue este último, por su mejor ajuste. En términos relativos, el Δ de los AIC entre estos modelos, en relación al tomado como óptimo, se resume en el Cuadro 5.4

Cuadro 5.2: Valores del Δ AIC para modelos LM en escenario VD

| Modelo | Δ AIC |
|---|---|
| 1 | 0.0163 |
| 2 | 0.0078 |
| 3 | 0 |

Los dos modelos GAM analizados, como se observa en la Figura 5.2, no presentan mayores diferencias a nivel gráfico entre el modelo con base *thin plate regression splines* y el que utilizó *cubic regression*, más según AIC el primero modelo presentó un 482918.38 en tanto el segundo un 481261.78, seleccionándose este último a pesar que la diferencia fuera mínima.

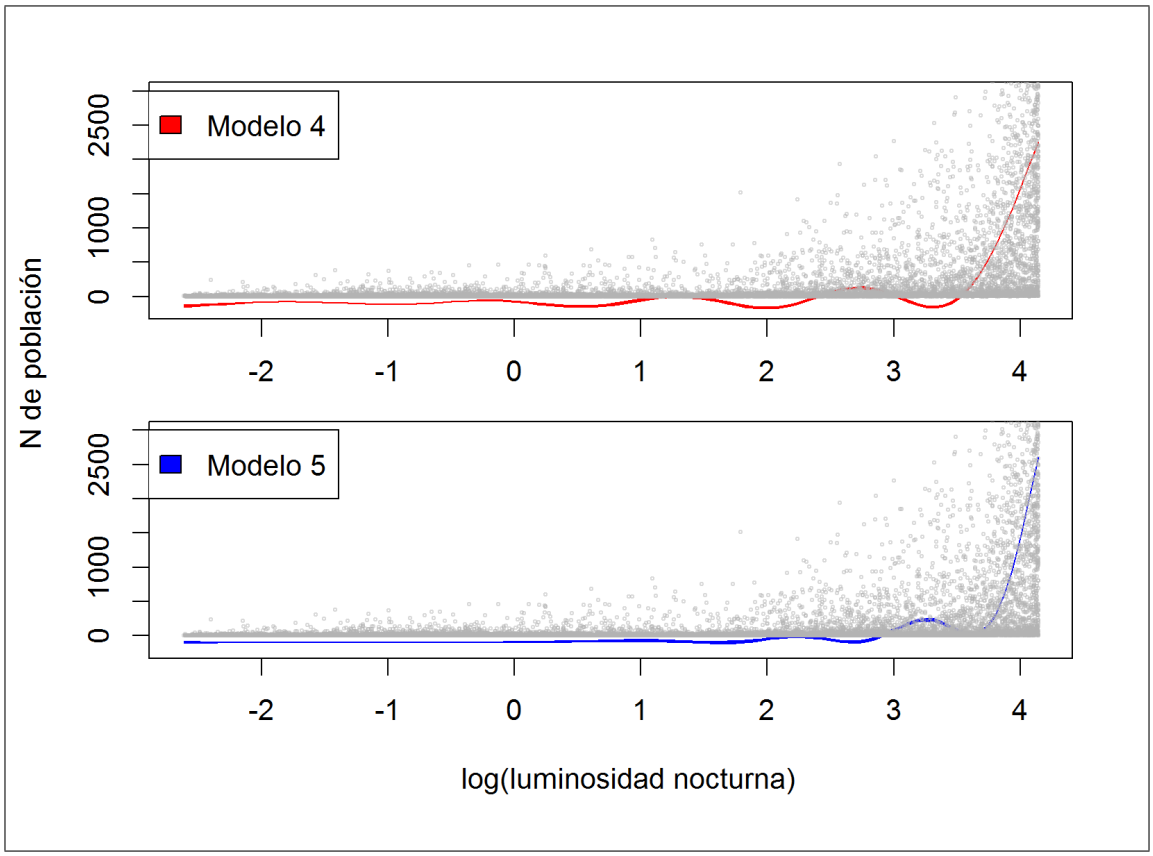


**Figura 5.2:** Evaluación modelos GAM de luminosidad y N poblacional. Fuente: Elaboración propia en base a procesamiento de imágenes satelitales OLS y marco censal 2011 (INE, 2014)

Para observar la diferencia entre ambos modelos GAM, se observó en términos relativos al modelo 5, el $\Delta$ de los AIC entre estos modelos, como se muestra en el Cuadro 5.3.

Cuadro 5.3: Valores del $\Delta$ AIC para modelos GAM en escenario VS

| Modelo | $\Delta$ AIC |
|---|---|
| 4 | 0.0034 |
| 5 | 0 |

Pasando ahora a los tres modelos CART considerados, vale recordar que corresponden a tres podas del árbol de mayor desagregación, por lo que la estructura de ramas presentó cada vez más densa como muestra la Figura 5.3 de los 3 modelos.

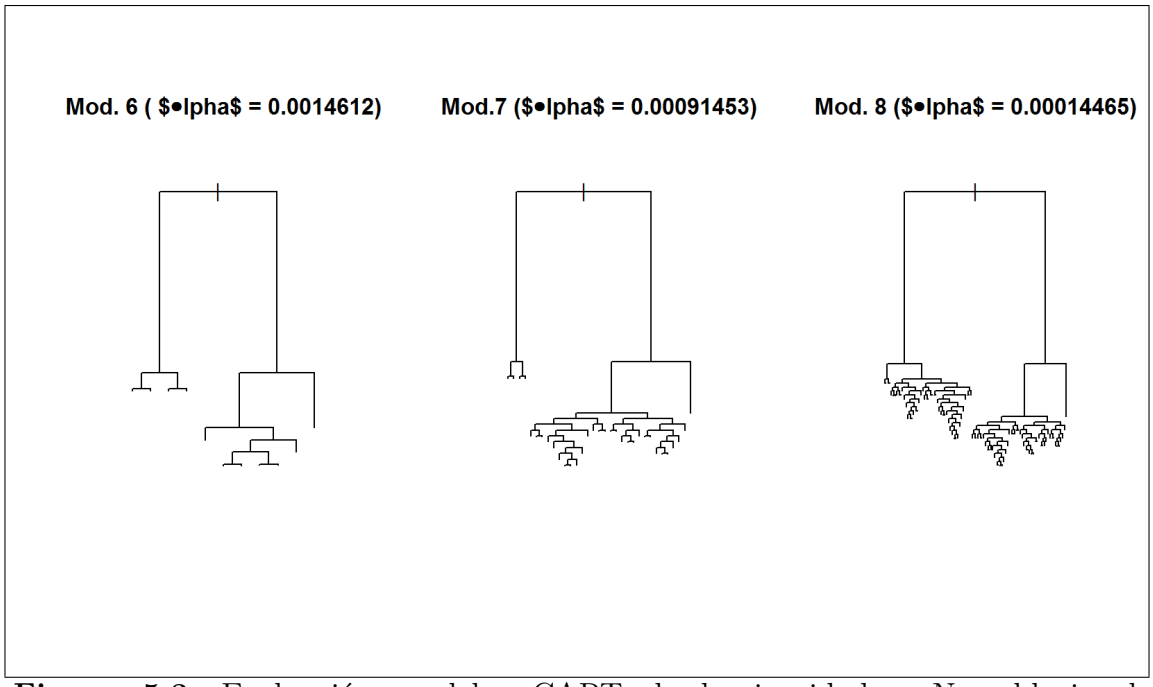


**Figura 5.3:** Evaluación modelos CART de luminosidad y N poblacional. Fuente: Elaboración propia en base a procesamiento de imágenes satelitales OLS y marco censal 2011 (INE, 2014)

No se muestra el etiquetado de los árboles, dado que solo se pretende ejemplificar la complejidad de los mismos, evidenciada en la centidad de aperturas de brotes, aunque es posible una interpretación desde los valores de corte, no se presenta aquí. Pero cuando evaluamos los tres modelos por su varianza explicada, encontramos que el modelo 6 presentó un 55.93 %, mientras que el modelo 7 un 58.59 %, y el modelo

8 un 61.07 %. A partir de este estimador, es que se seleccionó el modelo 8 como el óptimo dentro de este grupo.

Es así que, pasando a la evaluación entre los dos modelos Bagging antes definidos, donde lo estocástico se presenta en forma de remuestras bootstrap, y que al momento de una primera evaluación según sus errores *out of bag* el modelo con 25 iteraciones presentó un error promedio de 417.77 personas, mientras que el modelo que iteró 100 veces lo redujo a 414.81 individuos, y aunque esta diferencia parezca reducida, se tomó este segundo modelo para la evaluación intermodelos dada la escasa diferencia en tiempo de computo entre los modelos.

De modo similar, pero haciendo la *caja* un poco más *negra* es que se evaluó dos modelos RF, uno con 200 iteraciones, y otro de 500, presentada en la Figura 5.4 la distribución del error según número de iteraciones, que demuestra que la selección de 100 iteraciones es razonable, por más que se podría haber computado más, optamos por tomar el segundo como modelo "ganador" entre ambos, pues aunque tenga un rango mayor de error *oob* que el primero, presenta un comportamiento asintótico más suavizado.

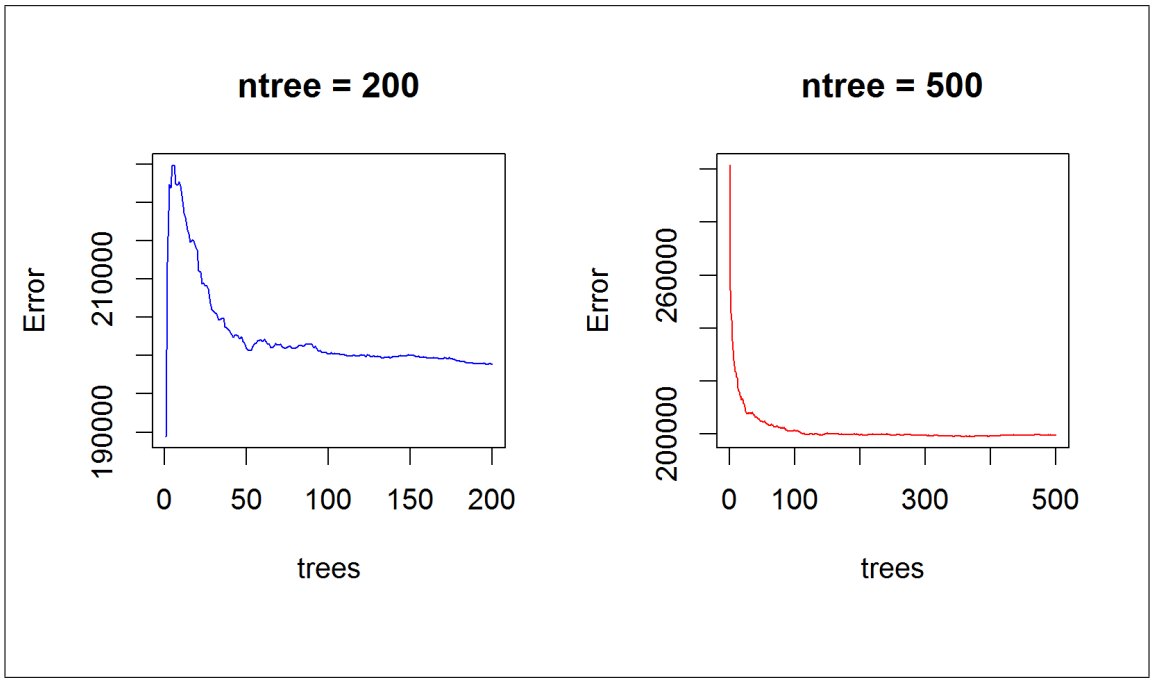


**Figura 5.4:** Evaluación modelos Random Forest de luminosidad y N poblacional. Fuente: Elaboración propia en base a procesamiento de imágenes satelitales OLS y marco censal 2011 (INE, 2014)

## 5.4. Resultados Variable Dasimétrica (VD)

El primer grupo correspondiente a modelos lineales, traen consigo el primer resultado, en tanto presentan una relación lineal entre las variables de luminosidad y cantidad poblacional evaluadas. Como muestra la Figura 5.5, entre los modelos LM, a mayor grado el ajuste mejora.

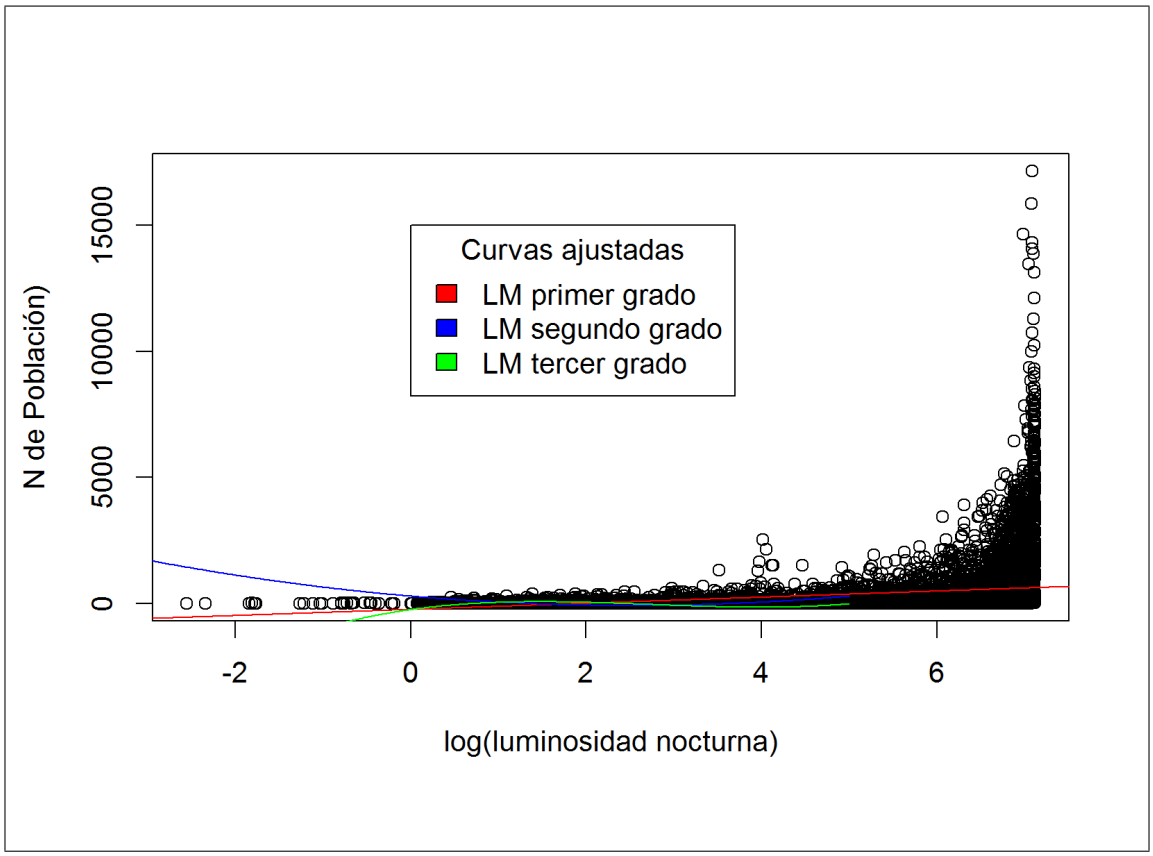


**Figura 5.5:** Ajuste de consumo dasimetrizado por luminosidad y N poblacional en modelos LM. Fuente: Elaboración propia en base a procesamiento de imágenes satelitales OLS, consumo residencial (UTE) y marco censal 2011 (INE, 2014)

En cuanto a sus resultados en la comparación por AIC, el modelo 1 arrojó un AIC de 532112.75, mientras que el modelo 2 (segundo grado) obtuvo un AIC de 527318.87, finalmente el modelo 3 configuró un AIC de 523079.33; por tanto, el modelo LM que se utilizó fue este último, por su mejor ajuste. En términos relativos el $\Delta$ de los AIC entre estos modelos en relación al tomado como óptimo se resume en el Cuadro 5.4.

Cuadro 5.4: Valores del Δ AIC para modelos LM en escenario VD

| Modelo | Δ AIC |
|---|---|
| 1 | 0.0173 |
| 2 | 0.0081 |
| 3 | 0 |

Por su parte, los dos modelos GAM considerados, presentaron la siguiente distribución en relación a los datos en la Figura 5.6. Como se observa, no hay mayores diferencias a nivel gráfico entre el modelo con base *thin plate regression splines* y el de *cubic regression*, de todos modos, según AIC el primero presentó un 518193.38 mientras que el segundo un 519823.28, quedándo este último como óptimo por más que la diferencia sea mínima.

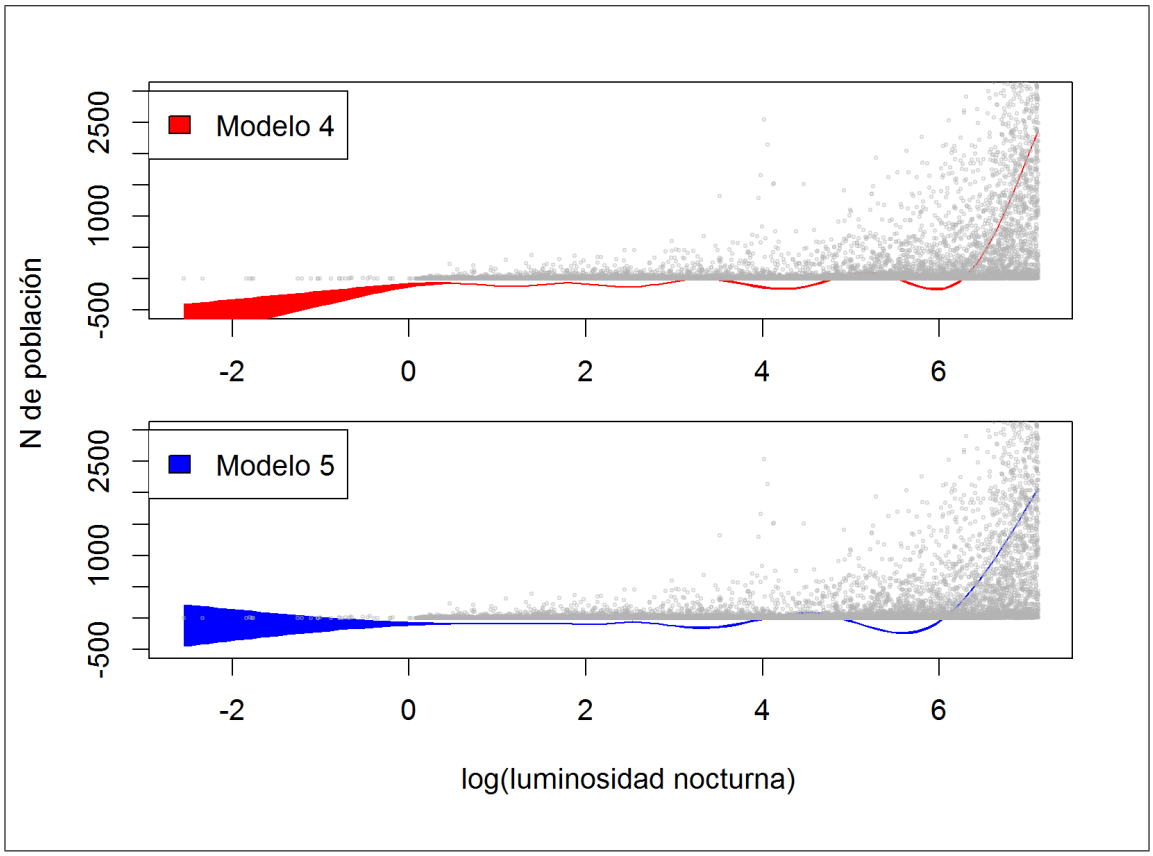


**Figura 5.6:** Evaluación modelos GAM de consumo dasimetrizado por luminosidad y N poblacional. Fuente: Elaboración propia en base a procesamiento de imágenes satelitales OLS, consumo residencial (UTE) y marco censal 2011 (INE, 2014)

Veamos ahora en el Cuadro 5.5, el Δ de los AIC entre estos modelos, para se observa la diferencia entre ambos modelos GAM en términos relativos al modelo 5, ya que según indica la Figura 5.6 fue el que se tomó como representante de este tipo de modelos.

Cuadro 5.5: Valores del Δ AIC para modelos GAM en escenario VD

| Modelo | Δ AIC |
|---|---|
| 4 | -0.0031 |
| 5 | 0 |

Entre los tres modelos CART considerados, podas del árbol de mayor desagregación, se muestra la estructura de ramas cada vez más densa, típica de esta construcción de modelos, presentándose en la Figura 5.7 los 3 modelos:

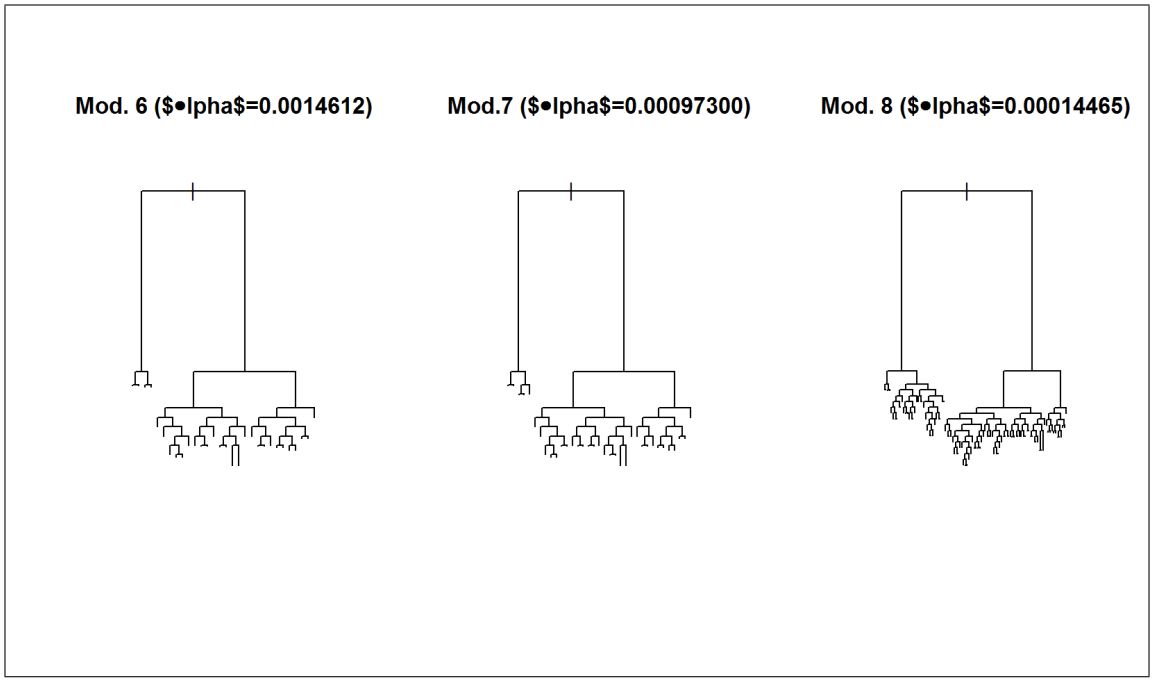


**Figura 5.7:** Evaluación modelos CART de consumo dasimetrizado por luminosidad y N poblacional. Fuente: Elaboración propia en base a procesamiento de imágenes satelitales OLS, consumo residencial (UTE) y marco censal 2011 (INE, 2014)

Del mismo modo que en el escenario anterior, no se muestra el etiquetado de los

árboles, pero los tres modelos al evaluarlos muestran que, el modelo 6 presentó un 61.94 %, en tanto el modelo 7 indicó un 62.55 %, y el modelo 8 un 65.22 % de su varianza explicada. Se seleccionó pues el modelo 8 como el óptimo dentro de este grupo.

En la evaluación entre los dos modelos definidos mediante el uso de Bagging, se observó que según sus errores *out of bag* el modelo con 25 iteraciones presentó un error promedio de 400.4687, y que el modelo que iteró 100 veces lo redujo a 395.7482. Esta diferencia puede parecer reducida, pero de todos modos se tomó este segundo modelo como el óptimo entre ambos.

Al evaluar los dos modelos RF, uno con 200 iteraciones, y otro de 500, se obtuvo que, como muestra la Figura 5.8, la distribución del error según número de iteraciones muestra que la selección de 100 iteraciones es razonable, por más que se podrían computar más.

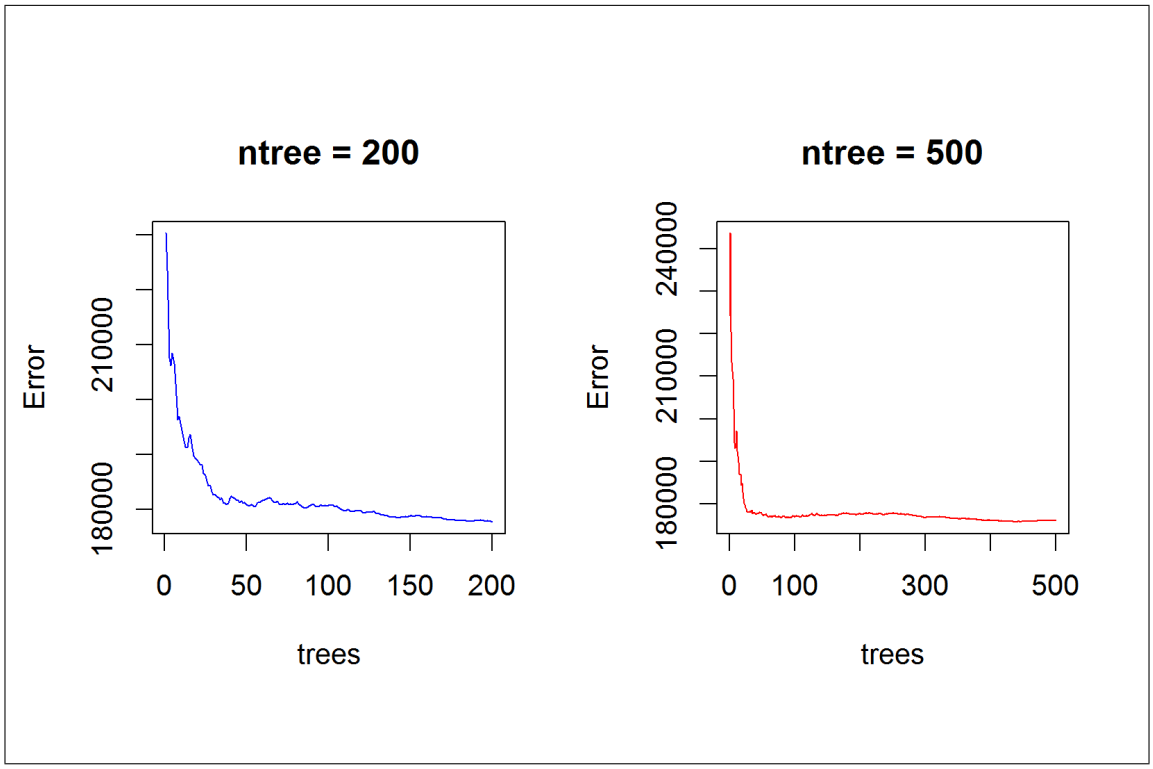


**Figura 5.8:** Evaluación modelos Random Forest de consumo dasimetrizado por luminosidad y N poblacional.
Fuente: Elaboración propia en base a procesamiento de imágenes satelitales OLS, consumo residencial (UTE) y marco censal 2011 (INE, 2014)

De todos modos, para optar entre ellos, y dado que la capacidad de computo no es un problema en estos tiempos, se tomó el modelo RF de 500 iteraciones, dado que

reduce el error más rápidamente.

## 5.5. Elección de modelos

Tras evaluar los 12 modelos de las 5 categorías indicadas, y habiendo definido al interior de cada grupo el representante óptimo para la estimación, se estimó para graficar los resultados de 20 iteraciones de cada modelo sobre muestras de aprendizaje y test (2/3 y 1/3 respectivamente).

Por más que se calcularon los errores de estimación de cada modelo, a nivel de muestras de aprendizaje y sobre muestra de test, se presenta en los siguientes apartados para cada escenario, la performance en las 20 iteraciones del porcentaje de varianza explicada por cada modelo, sobre muestra aprendizaje y de test, ya que esto es una cantidad más interpretable para evaluar los modelos.

### 5.5.1. Escenario Variable Sintomática

Para este primer escenario, la Figura 5.9 presenta la performance del mejor representante de cada tipo de modelo evaluado, correspondiendo: 1 - LM; 2 - GAM; 3 - CART; 4 - Bagging; y 5 - RF.

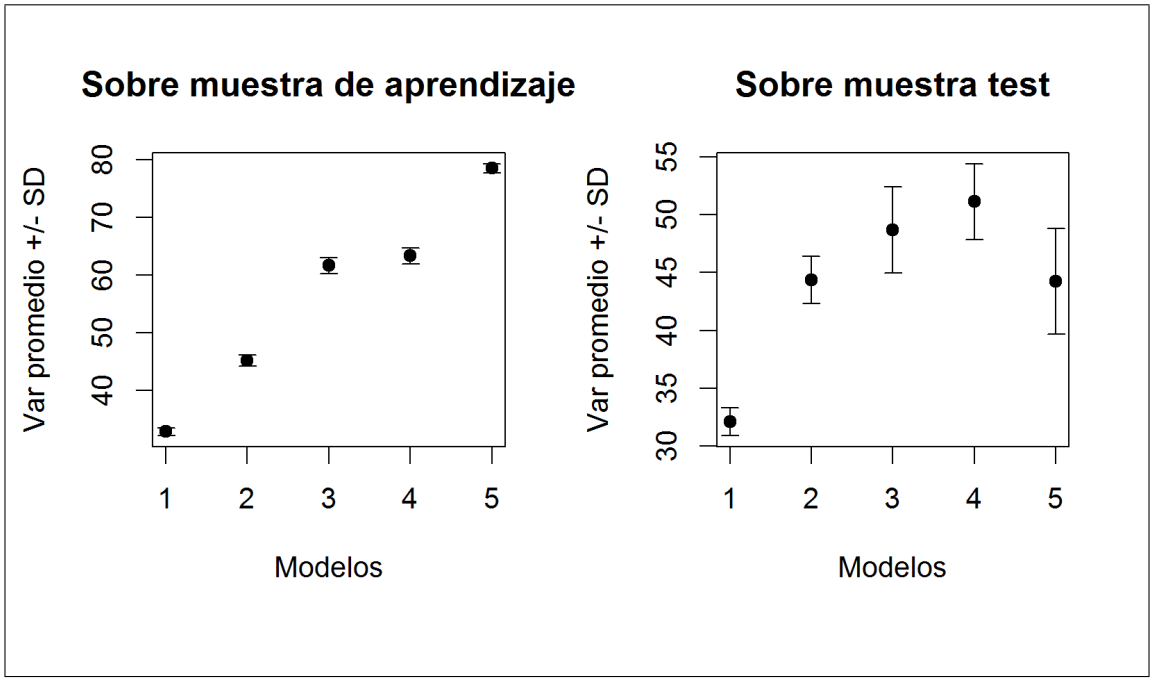


**Figura 5.9:** Comparación del % de varianza explicada promedio y su SD de los 5 modelos del escenario VS. Fuente: Elaboración propia en base a procesamiento

A partir de estos resultados, es claro que los modelos LM no son los óptimos para realizar las estimaciones según la relación entre la luminosidad nocturna y la cantidad de población asociada a nivel de píxel. De todos modos, entre los otros 4 modelos, es necesario realizar algunas decisiones metodológicas, en cuanto, el modelo CART en su modo podado nos permite interpretar los cortes de la regresión, por más que sea el de menor varianza explicada entre las muestras de aprendizaje y de test.

Por su parte, el modelo Bagging, aunque no mejora mucho su performance en relación al CART en las muestras de aprendizaje, en las test logra un adecuado 51.13 % de promedio de varianza explicada; y a pesar de que el modelo RF presente un 78.53 % de promedio de varianza explicada en las muestras, sobre las muestras de test solo logra un 44.24 %.

En definitiva, para este ejercicio, se tomó como el mejor modelo entre los evaluados al Bagging, dado que el objetivo es estimar con la mayor precisión posible a partir de imágenes satelitales de luminosidad nocturna la cantidad poblacional, y las variaciones esperadas son mejor representadas por las estimaciones en muestras test.

### 5.5.2. Escenario Variable Dasimétrica

Del mismo modo que en el escenario anterior, en la Figura 5.10, se presenta la performance de los modelos seleccionados en el marco del escenario de uso de la luminosidad nocturna de cada tipo de los evaluados: 1 - LM; 2 - GAM; 3 - CART; 4 - Bagging; y 5 - RF. Donde se puede observar nuevamente, que los modelos LM y GAM no compiten entre los óptimos para realizar estas estimaciones, por lo que, al pasar a evaluar entre los otros 3 modelos, se observó que el modelo CART en su modo podado nos permite interpretar los cortes de la regresión, por más que sea el de menor varianza explicada entre las muestras de aprendizaje y el segundo menor en muestras test.

El modelo Bagging, vuelve a no mejorar mucho su performance en relación al CART en las muestras de aprendizaje, pero muestra en las de test un interesante 51.13 % de promedio de varianza explicada; mientras que el modelo RF, presentó un 78.53 % de promedio de varianza explicada en las muestras, pero sobre las muestras de test solo logra un 44.24 %.

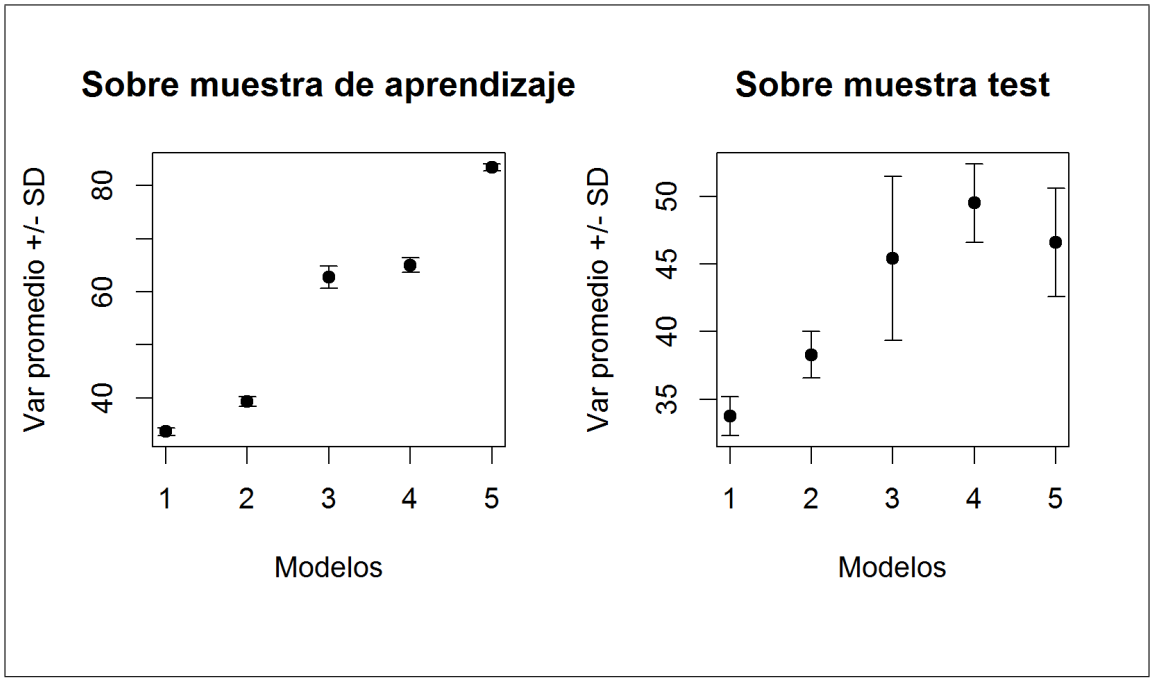


**Figura 5.10:** Comparación del % de varianza explicada promedio y su SD de los 5 modelos del escenario VD. Fuente: Elaboración propia en base a procesamiento

Es así que, en este escenario también se tomó como modelo óptimo entre los evaluados al Bagging, en virtud de que el objetivo es estimar con la mayor precisión posible la cantidad poblacional, y las variaciones esperadas son mejor representadas por las estimaciones en muestras test.

De todos modos, esto ha sido una primera aproximación, y dados los alentadores resultados de esta variable sintomática, es considerablemente alta la posibilidad de mejorar el modelo, utilizando censos y conteos anteriores, o considerando otras variables espacializables para complementar.

# 6. Resultados

## 6.1. Introducción

En este capítulo, se presentan las estimaciones de población para cada año intercensal (1996-2010) según cada escenario, y a partir de su modelo óptimo, de entre los evaluados en esta investigación. Luego se comparan estos resultados en relación con las estimaciones oficiales a nivel agregado por departamento, y finalmente con el censo de 1996 y el conteo poblacional de 2004 a nivel de píxeles.

En una primera instancia, se agregaron estas estimaciones por departamento para comparar los resultados de estas metodologías frente a las estimaciones publicadas por el INE, consideradas oficiales para el período. En una segunda instancia, se evaluaron las estimaciones hechas a nivel de píxel frente a los datos de población correspondientes al Censo de 1996 y el conteo de 2004, sorteando la población de cada segmento censal 2011, que no hayan sido modificado en su geometría al 2011.

## 6.2. Estimaciones nacionales

Tanto en el escenario de la luminosidad nocturna como variable sintomática de la cantidad de población, como en el escenario donde la luminosidad nocturna es utilizado como variable dasimétrica del consumo eléctrico residencial, usándose en conjunto como variable sintomática de la cantidad de población, el modelo considerado óptimo entre los evaluados, corresponde a los de Bagging.

Con las correspondientes imágenes de luminosidad nocturna, se estimó la población por píxel para cada año entre 1996-2010 para cada escenario, dando como resultados las siguientes curvas de población a nivel país. Ahora bien, se debe recordar que las estimaciones realizadas están únicamente ancladas a los datos del censo 2011, y como se puede observar en la Figura 6.1, en 2011 sería cuando mejor ajusta, pero está alejado en relación al total de población de 1996, y presenta un valle entre los años 2004 y 2008 que merece un análisis particular.

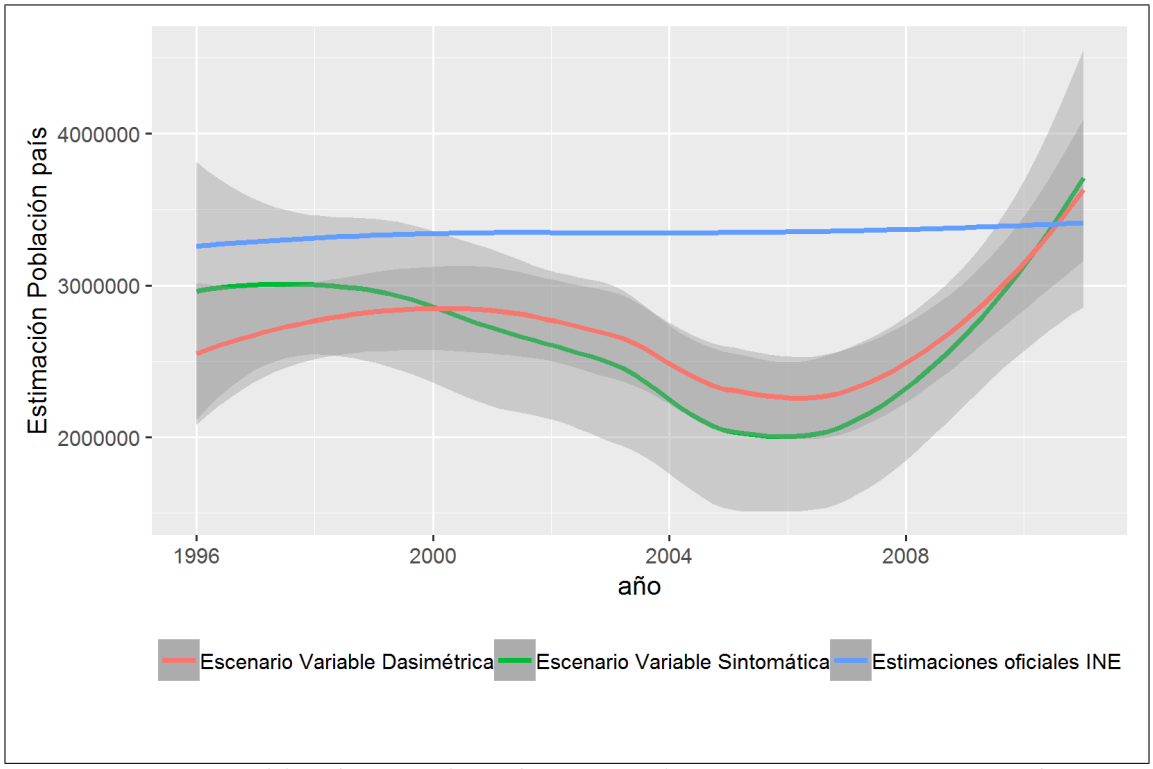


**Figura 6.1:** Población total país estimada con escenario sintomático y dasimétrico para todo el período junto a las estimaciones oficiales INE. Fuente: Elaboración propia en base a procesamiento de imágenes satelitales OLS, consumo residencial (UTE) y marco censal 2011 (INE, 2014)

Por un lado se debe considerar la variación en el stock poblacional y por otro los cambios en el patrón de consumo de energía eléctrica para iluminación residencial. Respecto al primer punto, sería de esperar que sí se incluyeran los datos del censo 1996 o del conteo 2004 a nuestros modelos, fuera de ajustarse mejor a los datos de la estimación oficial en esos años, la curva debería de reducir considerablemente el valle observado entre 2004 y 2008. Mientras que en el segundo punto, el cruce entre las curvas de estimación de los modelos en el año 2000, indica que el consumo residencial (incluido en la estimación VD) no cayó de igual forma que la luminosidad nocturna per se (único componente de estimación en el escenario VS).

En los gráficos de la Figura 6.2, se pueden observar los efectos de generar modelos alternativos, anclando en otros años con información (Censo 1996 y conteo 2004). Con cualquiera de los tres anclajes adicionales probados se optimiza el ajuste con los datos a 1996, de todos modos en todos los gráficos las curvas tienden a bajar en el 2000, manteniendo el valle en todos. Tan solo en el modelo anclado en 2004 y 2011 las estimaciones del INE permanecen dentro de las bandas de error de los modelos estimados, incluso sin cruzarse las curvas de estimación de los escenarios VS y VD.

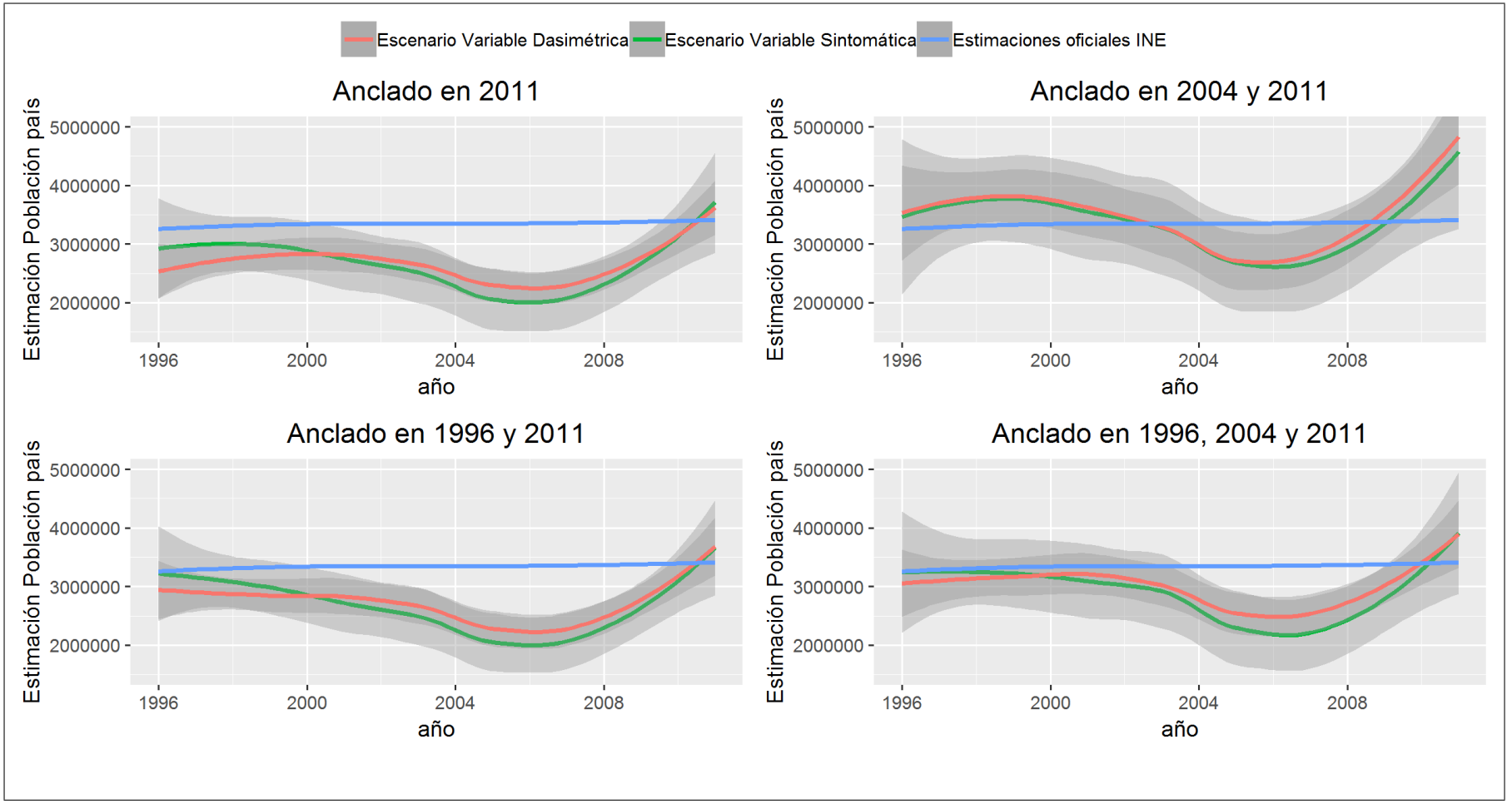


**Figura 6.2:** Comparación de predicciones anclando los modelos Bagging a los datos del censo 2011, y sus posibles combinaciones con el censo 1996 y el conteo 2004. Fuente: Elaboración propia en base a procesamiento de imágenes satelitales OLS, consumo residencial (UTE) y marco censal 1996, 2004 y 2011 (INE, 2014)

Tratar de explicar las causas de este comportamiento en detalle, implicaría un trabajo que excede los objetivos de esta tesis, pero se puede tener en cuenta varios aspectos que pueden arrojar luz al respecto. Primeramente, la crisis socio-económica que afecto a Uruguay en 2002, tiene sentido que haya influido tanto en el comportamiento emigratorio de la población como en la modificación del consumo eléctrico residencial, e incluso en la luminosidad nocturna propiamente dicha.

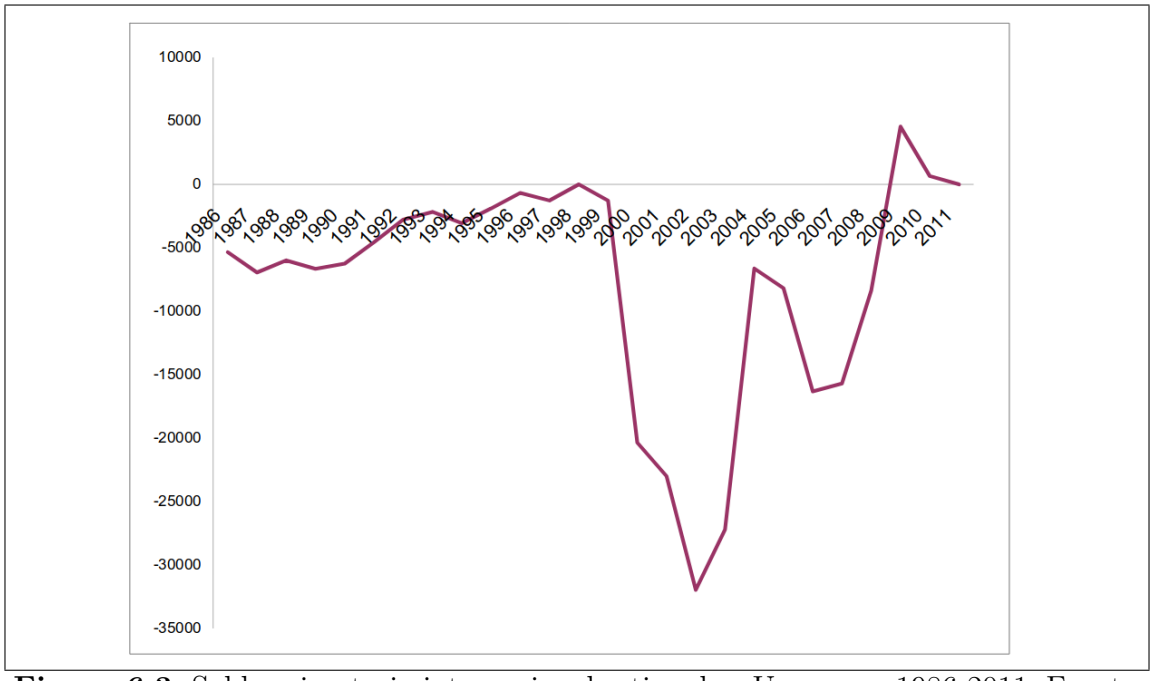


**Figura 6.3:** Saldo migratorio internacional estimado - Uruguay - 1986-2011. Fuente: INE (2014)

La Figura 6.3, puede en parte aportar a la explicación del descenso en las estimaciones de los modelos, más los rangos de error de las mismas dejan entrever las fragilidades de los modelos útilizados, fuera de como ya se mencionó, la situación de crisis económica con eje en el año 2002, también ha de haber cambiado en parte los consumos residenciales en busca de bajar gastos, o mismo los apagones zonificados de las luminarias públicas, también con fines de reducción de gastos pero desde la órbita estatal.

En las Figuras 6.4 y 6.5, se pueden observar ahora las estimaciones desagregadas por departamento para los dos escenarios, ordenados de mayor a menor población según censo 2011.

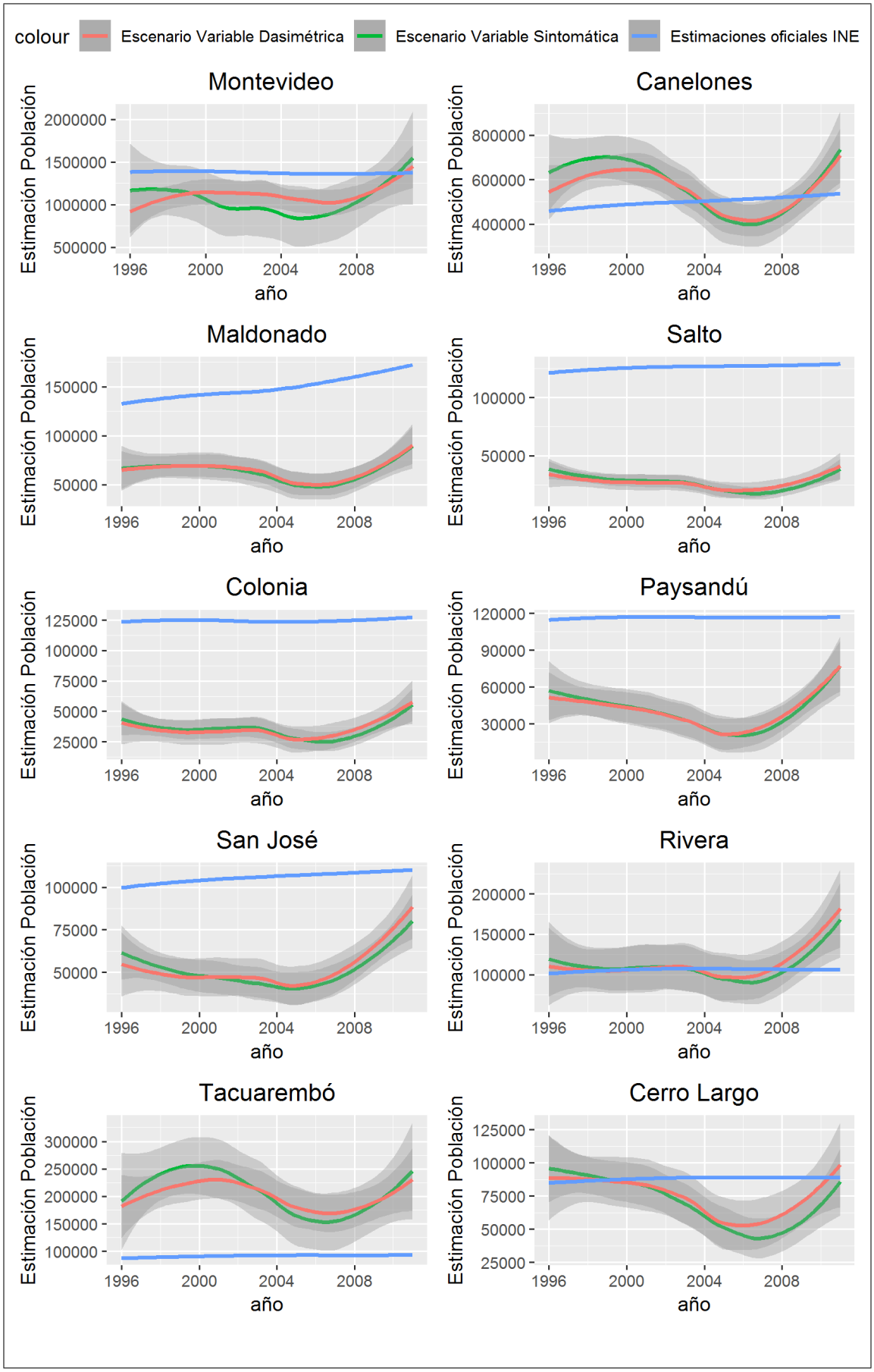


**Figura 6.4:** Población por departamento estimada con escenario sintomático y dasimétrico, junto a las estimaciones oficiales INE para todo el período. Parte 1 [49]

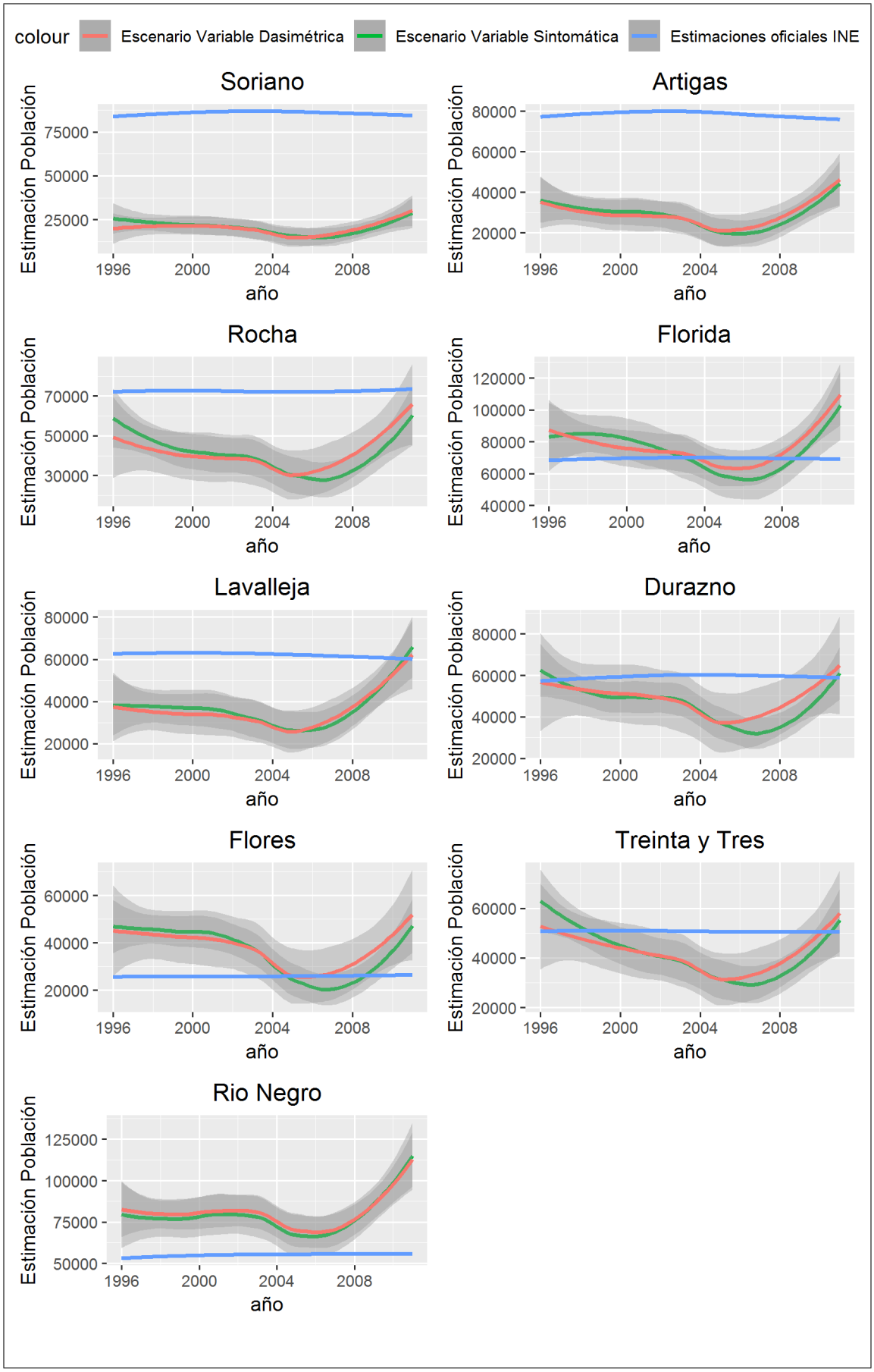


**Figura 6.5:** Población por departamento estimada con escenario sintomático y dasimétrico, junto a las estimaciones oficiales INE para todo el período. Parte 2 [50]

A partir de las Figuras 6.4 y 6.5, se pueden distinguir algunos comportamientos, Montevideo y Lavalleja ajustaron en 2011, pero devienen de subestimaciones en todo el período previo. Cerro Largo, Durazno y Treinta y tres, presentaron buen ajuste en 1996 y 2011, con marcado valle de subestimación entre 2004 y 2008.

Por su parte, Tacuarembó y Rio Negro fueron los únicos departamentos donde se sobreestimó la población para todo el período, mientras que Maldonado, Salto, Colonia, Paysandú, San José, Soriano, Artigas y Rocha fueron subestimados en todo el período. Canelones, Florida y Flores, ajustaron en el valle 2004-2008, subestimando en ambos extremos, Finalmente Rivera, fue el único caso donde ajustó muy bien para casi todo el período, despegando una sobreestimación para el tramo final 2008-2011.

De todos modos, estas estimaciones, cargan con posibles errores por saturación de los píxeles, y que a pesar de la movilidad de población de las zonas consolidadas de Montevideo, hacia zonas periféricas, esto no necesariamente se vería reflejado en los valores de luminosidad de los primeros en beneficio de los últimos, dado que pueden llegar a mantener la saturación del píxel sin esta población migrante, incluso interdepartamentalmente.

## 6.3. Evaluación de estimaciones

Para poder comparar los resultados de los escenarios indicados, fue necesario contar con indicadores resumen a nivel de las estimaciones, tanto para analizar las diferencias a pequeña escala entre las metodologías, como para corroborar la diferencia que hubo entre las agregaciones de las mismas a escala nacional, con las proyecciones oficiales en su última versión (INE, 2014). En este estudio se utilizó tres indicadores para comparar resultados: Error Porcentual Absoluto Medio (MAPE), Desviación Media Absoluta (MAD) y Raíz del Error Cuadrático Medio (RMSE).

El MAPE es el indicador sintético más utilizado para analizar comparativamente errores de estimación, que corresponde a la media aritmética que se ve afectada por los valores extremos y, a menudo exagera el error representado por la mayor parte de las observaciones en una previsión de población, siendo su definición:[9]

$$MAPE = \frac{100}{N} \sum \left| \frac{E_i - O_i}{O_i} \right|$$

Por su parte la MAD expresa precisión en las mismas unidades que los datos, lo que ayuda a conceptualizar la cantidad de error; se desestimó utilizar la Desviación

[9]Todas las definiciones de este apartado asumen a: O - Observaciones oficiales; E - estimaciones del modelo; y N - número de áreas.

Cuadrática Média (MSD) dado que es una medida comúnmente usada de la precisión de los valores de series de tiempo ajustadas, y los valores atípicos tienen un efecto mayor sobre MSD que en MAD. En concreto, se puede definir como:

$$MAD = \frac{\sum |O_i - E_i|}{N}$$

Siguiendo la recomendación de Legates y McCabe (1999), se calculó el error en las predicciones del modelo cuantificado en términos de las unidades de la variable calculada mediante la raíz del error cuadrático medio (RMSE). Este indicador es usado frecuentemente y su definición viene dada por:

$$RMSE = \sqrt{\frac{\sum |O_i - E_i|^2}{N-1}}$$

## 6.4. Estimaciones por departamentos 1996-2010

El comportamiento de los modelos óptimos de los dos escenarios, fue evaluado con los indicadores antes mencionados, con lo que se obtuvo en relación a las estimaciones del INE (2014) para el período 1996-2010 a nivel de departamentos los siguientes resultados.

Según los errores porcentuales absolutos medios (MAPE) el escenario sintomático presenta menor error promedio como se puede observar en la Figura 6.6, aunque si se considera el outlier en el escenario VS correspondiente al departamento 18 (Tacuarembó) presenta una media de 50.07 % de error, y el escenario dasimétrico 48.26 % de error.

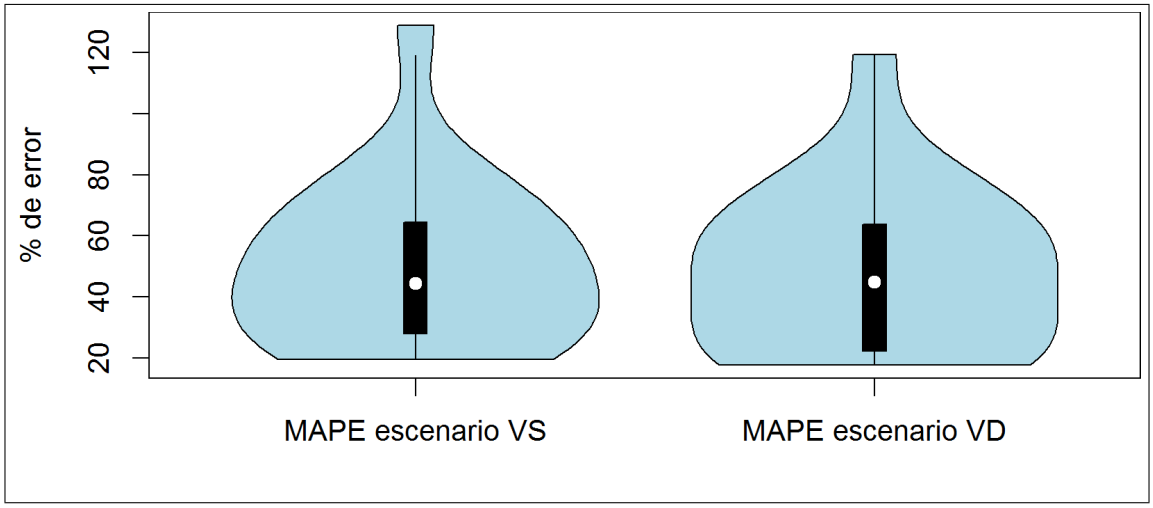


**Figura 6.6:** Evaluación del MAPE entre escenario sintomático y dasimétrico según estimaciones por departamento para todo el período.

Estos errores, pueden ser asociados cómo en los encontrados en las estimaciones de población de Dória (2015), que presentaron errores muchas veces por encima del 20 % y 30 %. Asociándose estos errores a factores locales de algunas manchas de luz o, a continuación, a las limitaciones inherentes al sistema/sensor, cómo la saturación del brillo de los píxeles en los centros urbanos más consolidados.

Por su parte, la Desviación Media Absoluta (MAD), presenta dos comportamientos diferenciables, si observamos los valores positivos y negativos de la Figura 6.7, es posible considerar que ambos modelos estiman por exceso la población.

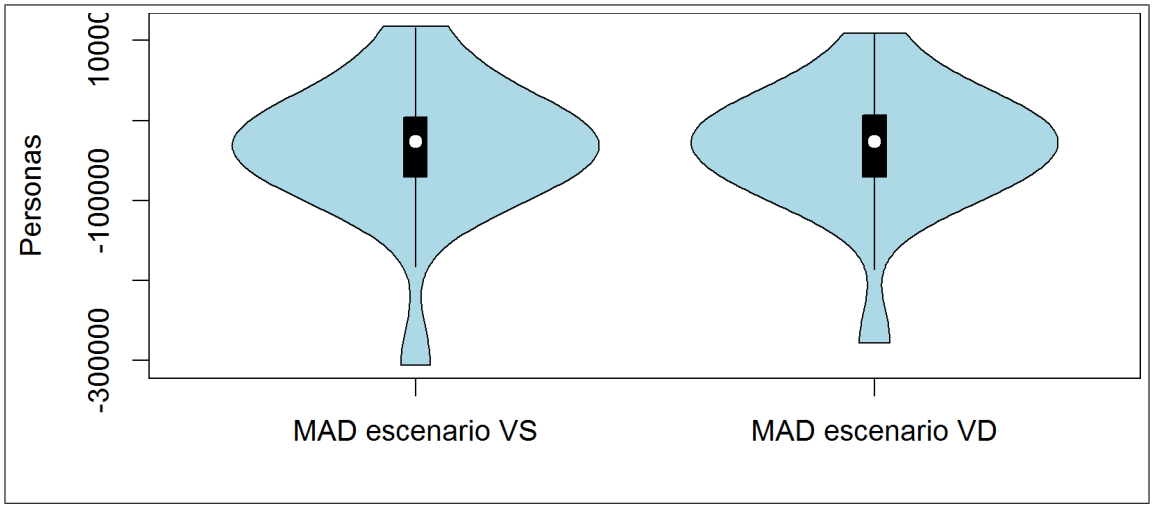


**Figura 6.7:** Evaluación del MAD entre escenario sintomático y dasimétrico según estimaciones por departamento para todo el período.

Por otro lado, al hacer foco en el comportamiento distintivo del departamento 1 (Montevideo), por ser la capital del país y contener aproximadamente la mitad de la población nacional, se puede indicar que el escenario de variable sintomática ajusta mejor en departamentos con mayor población.

Finalemente, mediante la raíz del error cuadrático medio (RMSE), que también se expresa en población, pero que al ser cuadrático evita que los errores opuestos se cancelen mutuamente, se puede observar que el escenario de variable dasimétrica tiene, aunque poco significativa, una mejor performace.

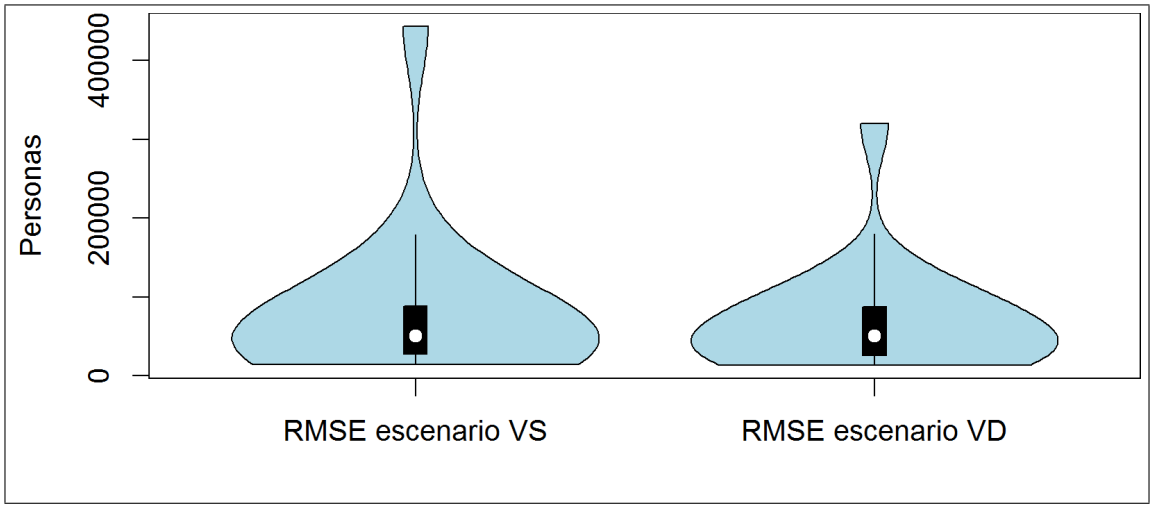


**Figura 6.8:** Evaluación del RMSE entre escenario sintomático y dasimétrico según estimaciones por departamento para todo el período.

## 6.5. Estimaciones por píxeles 1996 y 2004.

En este apartado, manteniendo los modelos Bagging tomados como los mejores estimadores, se evaluó en ambos escenarios el comportamiento de las estimaciones a escala de píxeles, evaluándolas frente a los datos de población correspondientes al Censo de 1996 y el conteo de 2004 sorteando la población de cada segmento censal 2011, que no hayan sido modificado en su geometría (resegmentados) entre ambos censos.

Recordemos que en 2004 se rezonificó pero no se resegmentó, y que en el pre censo 2011 además de efectuar las resegmentaciones pendientes desde 2004, se actualizaron las zonas amanzanadas, que a su vez, modificaron los límites de los segmentos para el nuevo censo. Todo esto presentó una dificultad, ya que a pesar de haber obtenido la tabla relacional, que indica que zonas censales de 2004 se dividían y/o unían en relación a lo disponibilizado como zonas censales a 2011, esta tabla no acompasaba las modificaciones geométricas estrictamente, y los problemas topológicos de la capa de segmentos. Por tanto, esta evaluación se hará exclusivamente para los 3.745 segmentos que no han variado en el pasaje de 3.958 segmentos en 1996 a los 4.313 de 2011.

Téngase en cuenta también, que las rezonificaciones y resegmentaciones intercensales, corresponden prinicipalmente a crecimientos urbanos, por lo que, las zonas periféricas, que normalmente acumulan una parte importante del crecimiento demográfico, no estarían siendo evaluadas en este apartado.

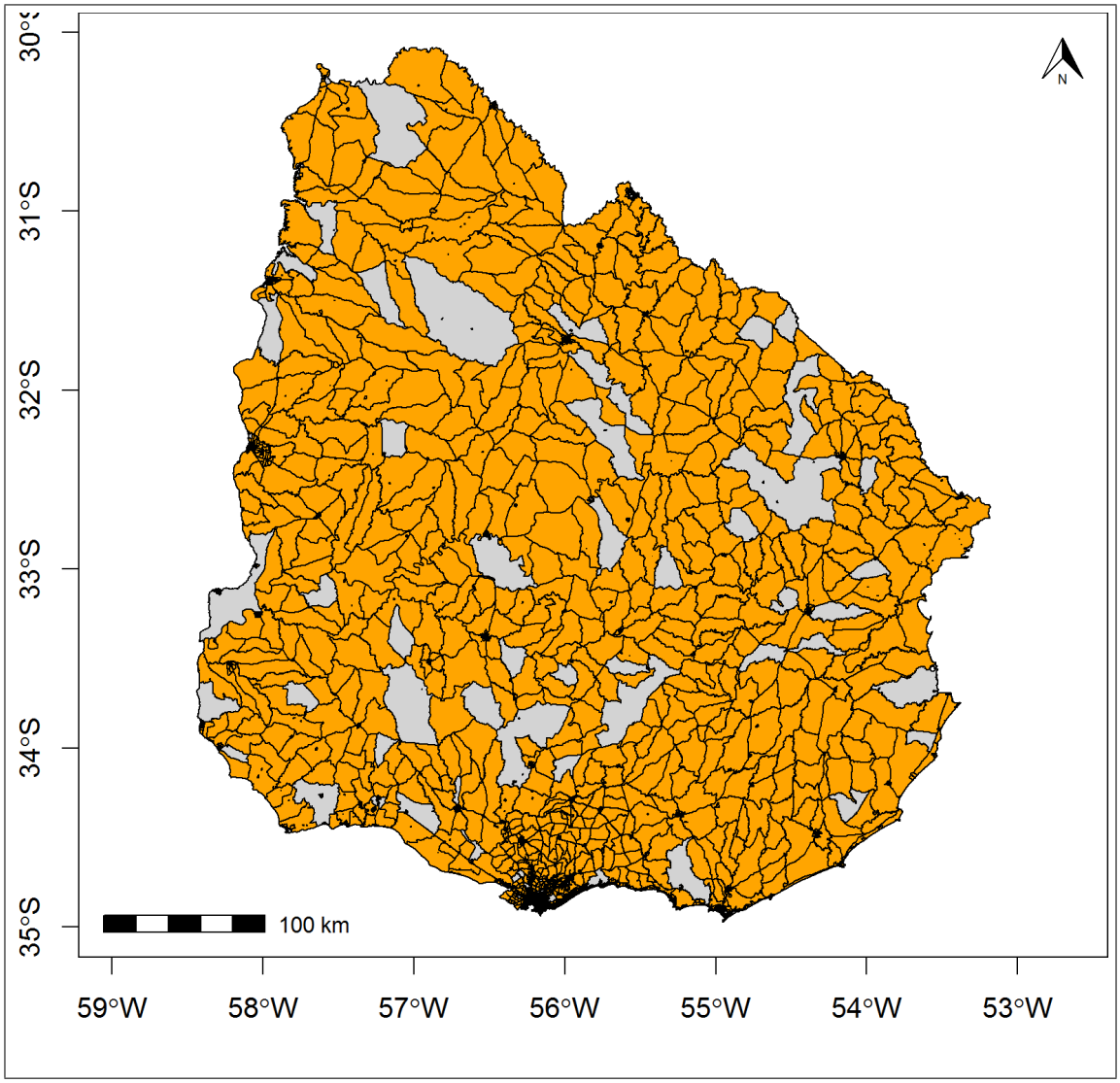


**Figura 6.9:** Distribución de los segmentos censales no modificados entre 1996 y 2011. Fuente: Elaboración propia en base a marco censal 1996 y 2011 (INE)

Al calcular el MAPE por píxel a nivel país, se puede observar un 420.45 % de error entre la estimación en el escenario VS y el censo 1996, de modo similar presenta un 388.96 % de error para el mismo año pero según el escenario VD. Por su parte, al evaluar los errores entre la estimaciones por píxeles a 2004 y el conteo del mismo año, se calculó un 436.21 % de error en el escenario VS, y un 397.88 % de error para el escenario VD.

Los errores MAD por píxel a nivel país, se estimó en -18.47 personas en relación al censo 1996, y en -27.32 personas para el mismo año pero según el escenario VD. Frente al conteo 2004, se calculó unos -21.82 individuos en el escenario VS, y unos -31.01 para el escenario VD.

En términos de RMSE, frente al censo 1996 presenta un error de 497.14 personas, y 538.23 para el mismo año pero según el escenario VD. En tanto, en el escenario VS frente al conteo 2004, se calculó un error de 487.98 personas, y unas 526.96 para el escenario VD.

Para desagregar estos errores, se observan las distribuciones por departamentos[10] de estas diferencias entre la población a 1996 según censo y las correspondientes estimaciones para ambos escenarios. Y por ejemplo, en la Figura 6.10 perteneciente al escenario VS, al igual que en la 6.11 del escenario VD, se observa como el comportamiento de Montevideo(1) presentó un rango de errores mucho mayor a que presentaron los otros departamentos, pero en todos los casos las medianas rondan el 0.

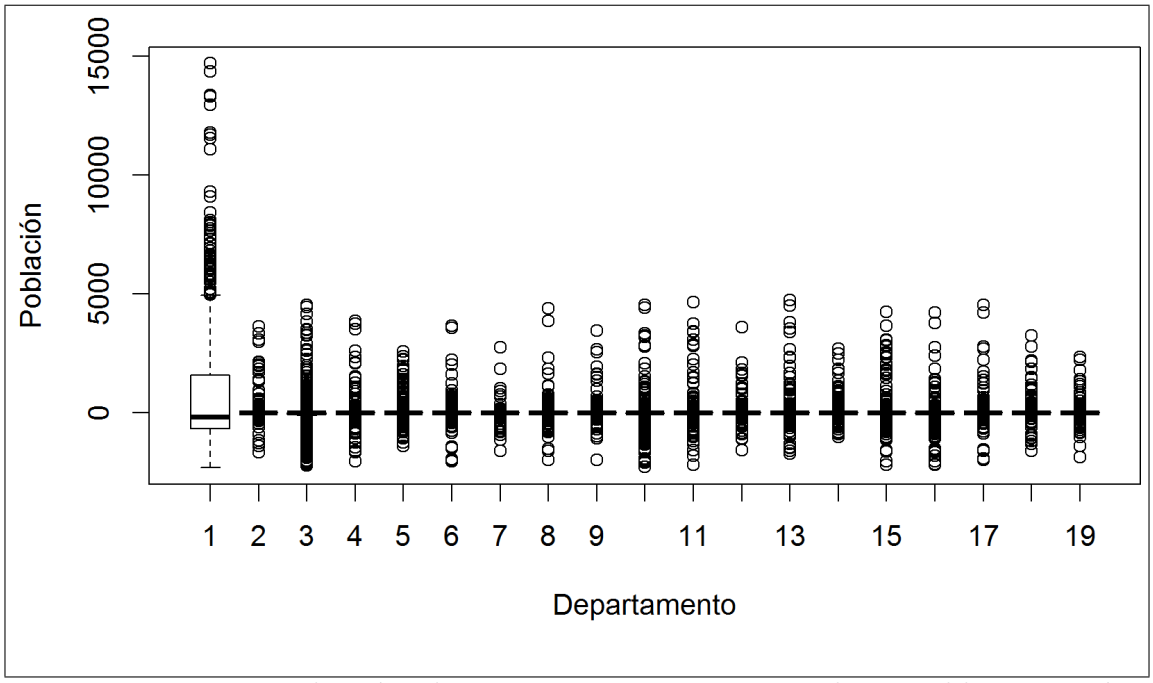


**Figura 6.10:** Distribución de errores entre escenario de variable sintomática según estimaciones a nivel de píxel para el censo 1996 por departamento. Fuente: Elaboración propia en base a procesamiento y marco censal 1996 y 2011 (INE)

[10]Nota: Para esta tesis, se ha utilizado la codiguera numérica de departamentos del INE, que Utiliza el 1 para Montevideo y luego enumera alfabeticamente al resto, o sea, 2-Artigas, 3-Canelones, 4-Cerro Largo, 5-Colonia, 6-Durazno, 7-Flores, 8-Florida, 9-Lavalleja, 10-Maldonado, 11-Paysandú, 12-Rio Negro, 13-Rivera, 14-Rocha, 15-Salto, 16-San José, 17-Soriano, 18-Tacuarembó y 19-Treinta y Tres

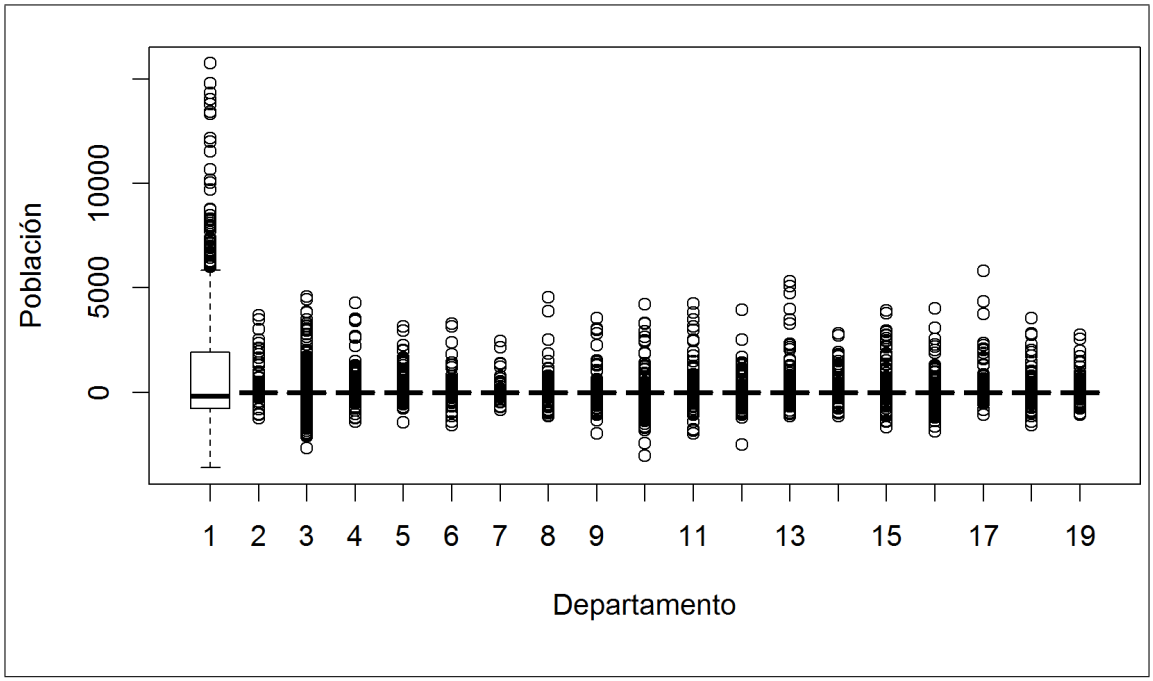


**Figura 6.11:** Distribución de errores entre escenario de variable dasimétrica según estimaciones a nivel de píxel para el censo 1996 por departamento. Fuente: Elaboración propia en base a procesamiento y marco censal 1996 y 2011 (INE)

Este último escenario, muestra de por sí un rango mayor en el caso de Montevideo (1) en relación al escenario anterior, de todos modos, la dispersión de los outliers en algunos departamentos da indicios de donde los errores presentan mínimos más dilatados, como son los casos de Canelones (3) y Maldonado (10), que son a su vez los siguientes en acumulación poblacional detrás de la capital.

Haciendo la misma comparativa entre las estimaciones y los resultados del conteo 2004, se observa un comportamiento similar.

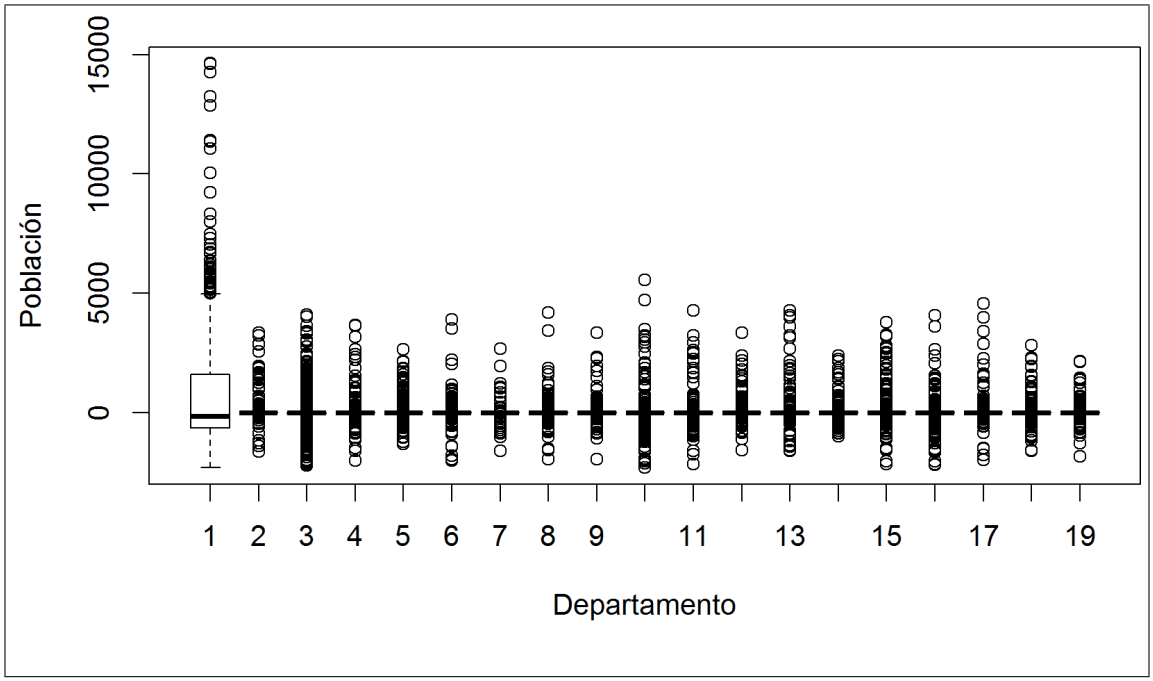


**Figura 6.12:** Distribución de errores entre escenario de variable sintomática según estimaciones a nivel de píxel para el censo 2004 por departamento. Fuente: Elaboración propia en base a procesamiento y marco censal 2004 y 2011 (INE)

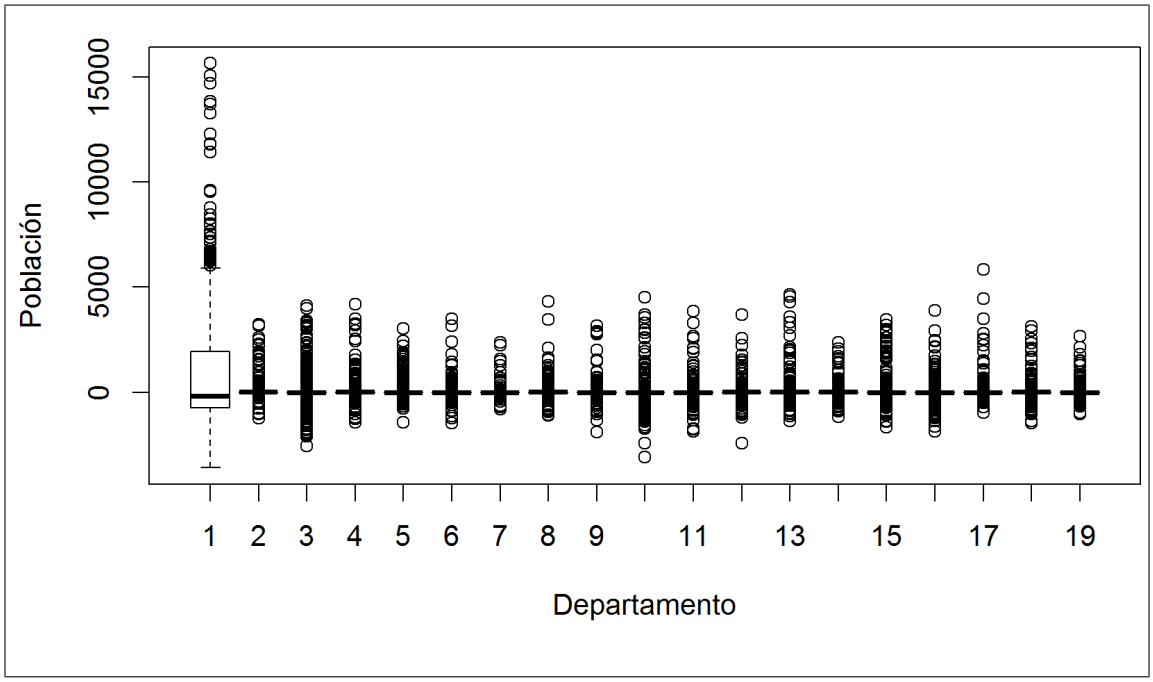


**Figura 6.13:** Distribución de errores entre escenario de variable dasimétrica según estimaciones a nivel de píxel para el censo 2004 por departamento. Fuente: Elaboración propia en base a procesamiento y marco censal 2004 y 2011 (INE)

Dados estos resultados, para arrojar luz respecto al ajuste de los modelos de ambos

escenarios, se agregaron los errores en relación al censo 1996 y conteo 2004 (a nivel de píxeles, de todos los segmentos no modificados entre 1996 y 2011) a nivel de departamento.

Al evaluar los MAPE de ambos escenarios con cada fuente, se puede observar los comportamientos diferenciales de las estimaciones por departamento, donde se puede destacar un 20.77 % promedio de error por departamento entre la estimación en el escenario VS y el censo 1996, de modo similar da un 20.13 % de error promedio para el mismo año pero según el escenario VD. Al evaluar los errores entre la estimación a 2004 y el conteo del mismo año, se calculó un 21.35 % de error promedio en el escenario VS, y un 20.62 % de error para el escenario VD en promedio.

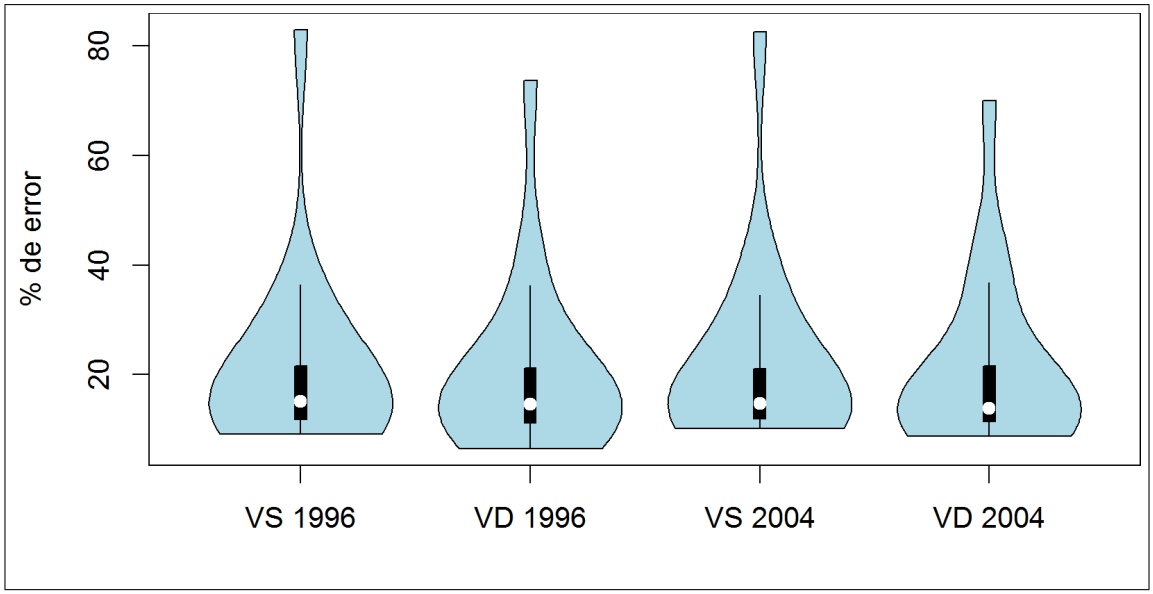


**Figura 6.14:** Evaluación del MAPE entre escenario sintomático y dasimétrico según estimaciones por píxel agregados por departamento para 1996 y 2004
Fuente: Elaboración propia en base a procesamiento y marco censal 1996, 2004 y 2011 (INE)

Por su parte, los MAD correspondientes a estos errores, presentados en población, corresponden a -52.23 en promedio por departamento en relación al censo 1996, y a unas -59.11 para el mismo año pero según el escenario VD. Frente al conteo 2004, que se calculó en unas -53.01 personas en promedio en el escenario VS, y unas -60 para el escenario VD de media.

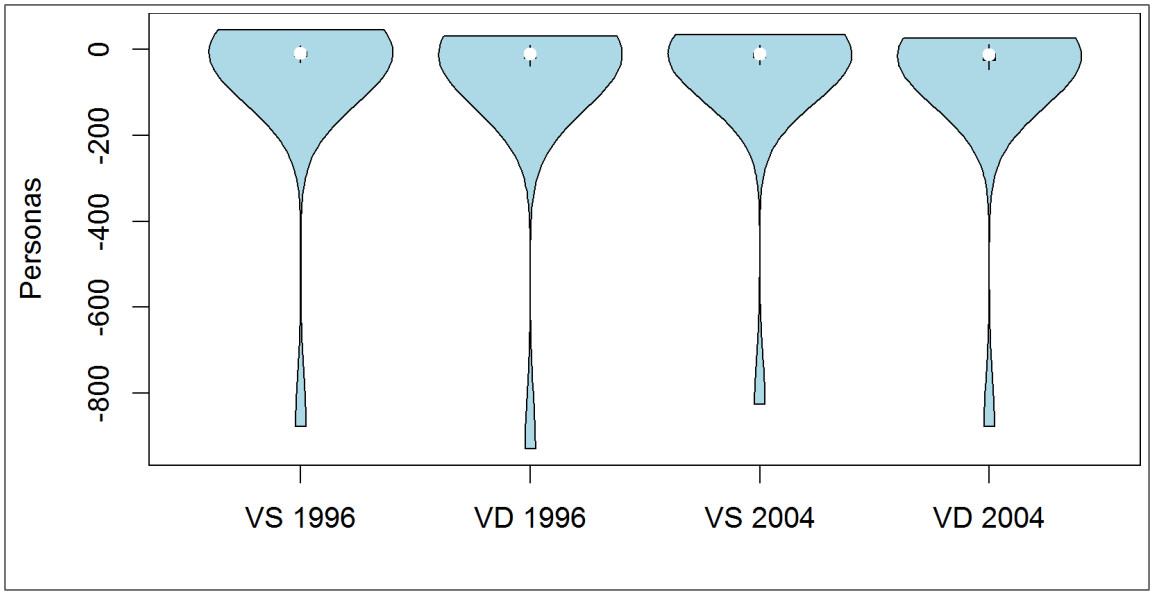


**Figura 6.15:** Evaluación del MAD entre escenario sintomático y dasimétrico según estimaciones por píxel agregados por departamento para 1996 y 2004
Fuente: Elaboración propia en base a procesamiento y marco censal 1996 y 2004 (INE)

Finalmente, los RMSE agregados por departamentos, que amplifican y penalizan con mayor fuerza aquellos errores de mayor magnitud, dan como resultado errores en el censo de 1996 que son en promedio en el escenario VS de 390.83 personas, en relación a las 407.96 del escenario VD promedio. Mientras que para el conteo de 2004, se calculó un error promedio de 387.61 personas, y unas 403.5 para el escenario VD.

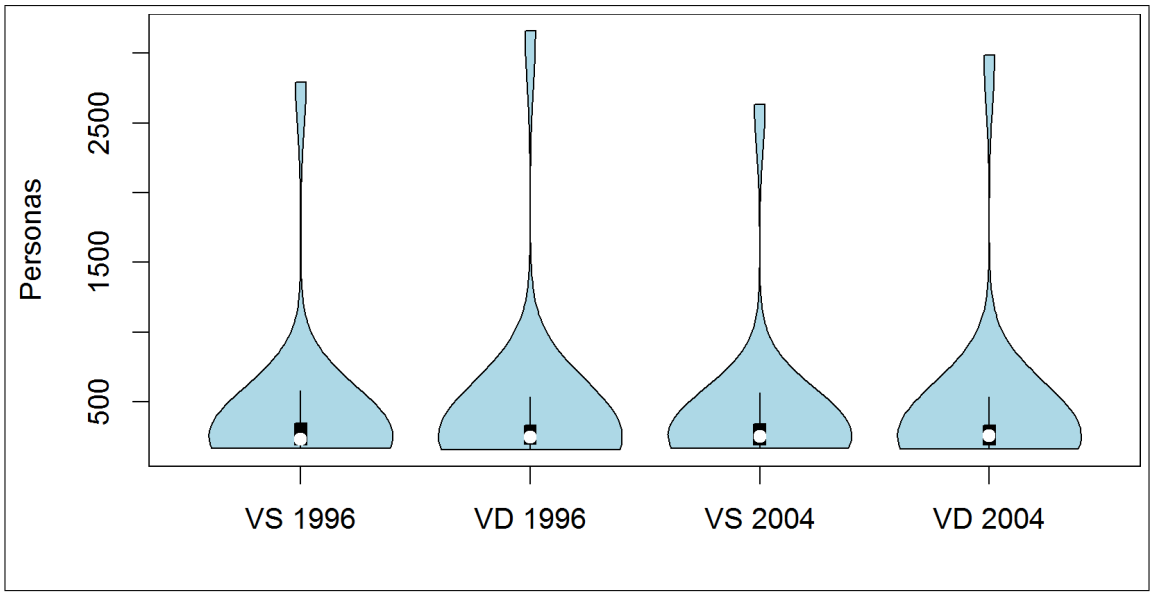


**Figura 6.16:** Evaluación del RMSE entre escenario sintomático y dasimétrico según estimaciones por píxel agregados por departamento para 1996 y 2004
Fuente: Elaboración propia en base a procesamiento y marco censal 1996 y 2004 (INE)

# 7. Conclusiones

## 7.1. Introducción

En este capítulo, se presentarán las conclusiones alcanzadas en tres sentidos, primeramente en relación a los resultados obtenidos en este estudio, en segundo término sobre los aprendizajes alcanzados en el proceso de investigación, y finalmente hacer incapié en los desafios pendientes que ha dejado esta tesis tanto a nivel personal como en esta rama específica del conocimiento.

## 7.2. Resultados

De lo analizado en el capítulo anterior, cabe destacar como conclusión que las fuentes de informaciòn utilizadas para la construcción de los modelos presentan un gran potencial, aunque las estimaciones hechas en ambos escenarios en las diferentes escalas temporales y espaciales no muestren una performance adecuada para la toma de decisiones con esta herramienta.

De todos modos, el análisis realizado ha permitido corroborar comportamientos diferenciales, entre las regiones menos y más pobladas, ya que fuera del efecto propio del tamaño del píxel (de 1 $km^2$), lo que más ha afectado las estimaciones ha sido la saturación de la luminosidad en los mismos. Esta saturación, que quiso ser controlada tanto por el uso de los valores logaritmizados (escenario VS), como por la dasimetrización de la información del consumo residencial eléctrico (escenario VD), mantuvo en ambos escenarios a las áreas urbanas con alta población con amplios errores a diferencia de áreas menos pobladas y sin saturación.

Haciendo una mirada longitudinal, cabe destacar el ajuste diferencial con los modelos anclados en las diferentes combinaciones de relevamientos poblacionales, y como aun quizás exageradamente las estimaciones de todos los modelos utilizados, muestran una curva descendente desde principios de los 2000 y un valle entre 2004 y 2008, señal que combina tanto la posible ola emigratoria que es normalmente dificil de estimar con el método de componentes, y a su vez, un cambio en el patrón de consumo energético también como posible estrategía post crisis del 2002, tanto a nivel residencial como estatal.

Vinculando los resultados de esta tesis, con los obtenidos por Cabrera (2011), se puede llegar a conclusiones similares a las arribadas por Barros (2017), que propone

que los datos provenientes de las imágenes satelitales nocturnas presentan menor potencial para la predicción de las variaciones poblacionales que los datos provenientes de los registros administrativos. Aunque, Barros (2017) también verificó que los diferenciales de cobertura y calidad de los registros administrativos y su variación en el tiempo, así como los cambios en la estructura de edad de la población, pueden impactar negativamente el desempeño de estos datos para predecir las variaciones poblacionales, haciendo que las estimaciones realizadas con el método tradicional (incrementos relativos) le haya sido el más cercano a los resultados del censo que aquellos provenientes de métodos que utilizan cualquier variable sintomática, siempre que haya datos de un conteo poblacional en el medio de la década.

## 7.3. Aprendizajes

La variedad de modelos evaluados en esta tesis, fuera de presentar el recorrido por el camino de complejización de los algoritmos para la evaluación de relaciones entre variables, permite mostrar el recorrido de aprendizaje entre el poder explicativo (Interpretabilidad) y el poder predictivo (Flexibilidad) de los modelos en busca de un equilibrio.

El haber trabajado con grillas regulares raster, dejó en evidencia los errores topológicos de las capas vectoriales del INE tanto en las geometrías 2011, y más acentuados aun en las 2004, publicadas ambas en la web oficial. Ante esta situación se sugiere a futuro pasar a utilizar grillas regulares para la desagregación espacial de la información estadística, tanto por su simplicidad topológica para procesamiento, como para evitar las rezonificaciones y resegmentaciones por las variaciones en el uso del suelo o categorización catastral.[11]

En otro orden, en ambos escenarios, se observa que las estimaciones tienden a subestimar la población, a la vez que muestran un comportamiento diferenciado entre los departamentos con alta cantidad de píxeles saturados (luminosidad = 63). Del mismo modo, asociándolo a los resultados de Dória (2015), y a la tendencia a

[11]Cabe destacar la experiencia llevada a cabo por el proyecto *Geostat* (ESSnet project Geostat)del Foro Europeo para la Geoestadística (EFGS), que desarrolla la generación de una malla formada por celdas de 1 km de lado, utilizando un mismo sistema de referencia espacial para la totalidad de Europa. Este proyecto realiza también tareas de estimación de la distribución por celdas de la población para un conjunto amplio de países en todo el continente, con celdas codificadas con un sistema estándar que sigue las indicaciones de la directiva Inspire. Y a nivel regional, recientemente el IBGE disponibilizó el producto *Grado Estadístico* (Geografia e Estatística (IBGE)., 2015), que representa un gran avance en la diseminación oficial de informaciones estadísticas provenientes de los censos (Bueno, 2014), en células equilaterales de 1 km en áreas rurales y 200 m en áreas urbanas.

sobre estimación que muestran los modelos hacia 2011, se puede considerar que en algunas regiones existan a 2011 diversos emprendimientos industriales que emiten luz durante la noche que antes no estaban, contribuyendo a un valor alto en la estimación, sin poseer población residente, por más que se hayan filtrado los píxeles sin población en el censo 2011.

Al observar los comportamientos de los errores de estimaciones, tanto en la evaluación frente a las estimaciones oficiales a nivel de departamentos, como al hacerlos frente a los datos del censo de 1996 y el conteo 2004, se podría concluir que las estimaciones realizadas en el escenario de variable dasimétrica presenta errores menores a los del escenrios de variable sintomática.

A partir de esto, es que debemos plantearnos desafíos pendientes, tanto en la obtención de herramientas de estimación más precisas, como en la búsqueda o construcción de nuevas y mejores fuentes de datos abiertas, y disponibles para el desarrollo de esta línea de investigación tanto a nivel local como para exportar metodologías a nivel global en pos de la posibilidad de estimaciones intercensales a pequeña escala de calidad.

## 7.4. Desafíos pendientes

Se considera que esta investigación amplía la puerta dentro de los estudios de población, tanto por incursionar a nivel nacional con fuentes no antes utilizadas en el marco de las estimaciones poblacionales, como en el amplio abanico de modelos planteados, que solo dan a descubrir la punta de iceberg de una caja de herramientas basta con aplicaciones cada día en más campos.

Al considerar abierta esta línea de trabajo sobre estimación y proyección de poblaciones en áreas pequeñas en Uruguay, es que se puede considerar como desafíos pendientes, el profundizar la apuesta y estar atentos a incluir nuevas variables a estos modelos, en pos de reducir los márgenes de error.

Asimismo se destaca de forma explícita, el formato de tesis mediante el uso de RMarkDown, que además de hacer transparente todo el trabajo realizado, al momento de evaluarla, permite replicar todo el análisis realizado, y actualizar todo el documento con solo mejorar los inputs. Por ejemplo, con las nuevas imágenes que reducen la resolución espacial, proceso que inició cuando el nuevo sensor Visible Infrared Imaging Radiometer Suite VIIRS fue colocado en órbita. Desde 2012, se cuenta con imágenes con mayor resolución que las anteriores, permitiendo obtener

resultados más precisos (Elvidge et al., 2013).

Finalmente cabe mencionar que esta tesis es parte del preámbulo y cimiento de mi propia tesis doctoral en curso, donde varios de los resultados y experiencias de implementación metodológica son aplicables, por más que la misma discurra por autómatas Celulares y Modelos Basados en Agentes, también para la estimación y proyección de población en áreas pequeñas.

# 8. Bibliografía

Amaral, S. et al. (2005) Estimating population and energy consumption in brazilian amazonia using dmsp night-time satellite data. *Computers, Environment and Urban Systems*. 29 (2), 179–195.

Barros, L. F. W. (2017) Potencialidades e desafios na utilização de registros administrativos e de imagens noturnas de satélite para realização de estimativas populacionais municipais intercensitárias no brasil. Tesis Doctorado en Población, Território y Estadísticas Públicas thesis. Rio de Janeiro: Escuela Nacional de Ciencias Estadísticas, ENCE.

Bay, G. (1998) El uso de variables sintomáticas en la estimación de la población de áreas menores. *Revista Notas de Población*. 67/68181–208. [online]. Disponible en: http://www.cepal.cl/publicaciones/xml/1/5431/ LCG2048_p7.pdf.

Bennett, M. M. y Smith, L. C. (2017) Advances in using multitemporal night-time lights satellite imagery to detect, estimate, and monitor socioeconomic dynamics. *Remote Sensing of Environment*. 192176–197.

Bharti, N. et al. (2011) Explaining Seasonal Fluctuations of Measles in Niger Using Nighttime Lights Imagery. *Science*. [Online] 334 (6061), 1424–1427. [online]. Disponible en: http://www.sciencemag.org/cgi/doi/10.1126/science.1210554 (Fecha de consulta 10 Julio 2018).

Breiman, L. (1996) Bagging predictors. *Machine Learning*. 26 (2), 123–140.

Breiman, L. (2001) Random forests. *Machine Learning*. 45 (1), 5–32.

Breiman, L. et al. (1984) *Clasification and regression trees*. Monterey, California, USA: Wadsworth, Inc.

Bryan, T. M. (2004) 'Population estimates.', en Jacob S. Siegel y David A Swanson (eds) *The methods and materials of demography*. 2nd edition San Diego, California: Elseiver Academic Press. págs 9–41.

Bueno, M. C. D. (2014) Grade estatística: Uma abordagem para ampliar o potencial analítico de dados censitários. Tesis Doctorado en Demografía thesis. Campinas: Instituto de Filosofia e Ciências Humanas da Universidade Estadual de Campinas.

Cabrera, M. (2011) *Estimación de Población en áreas menores con métodos que utilizan variables sintomáticas*. Montevideo: OPP – Comisión Sectorial de Pobla-

ción. [online]. Disponible en: http://www.opp.gub.uy/images/6._Estimacion_de_poblacion_en_areas_menores.pdf.

Calvo, J. J. y Prats, O. (1992) *Canelones : Proyecciones de la población 1985-2010 por sexo y grupo de edad.*

Calvo, J. J. y Rios, G. (1998) *Proyecciones de población y viviendas de la ciudad de salto, 1996 - 2025.*

CELADE (1981) *América Latina. Situación demográfica evaluada en 1980: Estimaciones (1960-1980) y Proyecciones (1980-2025).* Santiago de Chile: CELADE.

CELADE (1998) *Demografía ii.* México: Programa Latinoamericano de Actividades en Población, Carlos Welti Editor.

CELADE (1984) *Métodos para proyecciones demográficas.* San José, Costa Rica: CELADE.

Chávez Esquivel, E. (2001) Variables sintomáticas en las estimaciones poblacionales a nivel cantonal en costa rica. *Revista Notas de Población.* 7151–72. [online]. Disponible en: http://www.cepal.cl/publicaciones/ xml/3/7223/LCG2114_p3.pdf.

DGEC (1991) *Montevideo y resto urbano del país: Estimaciones y proyecciones de población por sexo y edad 1975-2025.* Montevideo, Uruguay: DGEC - CELADE.

Doll, C. (2010) 'Population detection profiles of DMSP-OLS night-time imagery by regions of the world.', en *Proceedings of the 30th Asia-Pacific Advanced Network Meeting.* 2010 Hanoi, Vietnam:.

Doll, C. N. et al. (2006) Mapping regional economic activity from night-time light satellite imagery. *Ecological Economics.* [Online] 57 (1), 75–92. [online]. Disponible en: http://linkinghub.elsevier.com/retrieve/pii/S0921800905001254 (Fecha de consulta 2 Mayo 2017).

Dória, V. E. M. (2015) Sensoriamento remoto de luzes noturnas para estimativas populacionais em escalas regional e local: Os casos do distrito florestal sustentável da br-163 (pa) e da região metropolitana de são paulo. Tesis Maestria en Sensoriamento Remoto thesis. São José dos Campos: Instituto Nacional de Pesquisas Espaciais (INPE).

Duchesne, L. (1988) *Proyecciones de Población por sexo y edad para áreas intermedias y menores; método de relación de cohortes.* CELADE.

Elvidge, C. D. y Sutton, K. E. Z., P. C.; Baugh (2011) National trends in satellite

observed lighting: 1992–2009. *Remote Sensing*. 318.

Elvidge, C. D. et al. (1997) Mapping city lights with nighttime data from the DMSP operational linescan system. *Photogrammetric Engineering & Remote Sensing*. 63 (6), 727–73.4.

Elvidge, C. D. et al. (2013) Why viirs data are superior to dmsp for mapping nighttime lights. *Proceedings of the Asia-Pacific Advanced Network*. 62–69.

Elvidge, C. D. et al. (2001) Night-time lights of the world: 1994-1995. *ISPRS Journal of Photogrammetry and Remote Sensing*. 56 (2), 81–99.

Elvidge, C. D. et al. (2009) A fifteen year record of global natural gas flaring derived from satellite data. *Energies*. 2595–622.

Feng-Chi, H. et al. (2013) *Exploring and estimating in-use steel stocks in civil engineering and buildings from night-time lights.* 34490–504.

Fisher, R. A. (1936) The use of multiple measurement in taxonomic problems. *Annals of Eugenics*. 7 (2), 179–188.

Gauss, C. (1809) *Theoria motus corporum coelestium in sectionibus conicis solem ambientium auctore carolo friderico gauss*. sumtibus Frid. Perthes et I. H. Besser. [online]. Disponible en: https://books.google.com.uy/books?id=VKhu8yPcat8C.

Gauss, C. (1821) Theory of the combination of observations which leads to the smallest errors. *Gauss Werke*. 41–93.

Geografia e Estatística (IBGE)., I. B. de (2015) *Grade estatística: Guia de utilização*. Rio de Janeiro:. [online]. Disponible en: ftp://geoftp.ibge.gov.br/malhas_digitais/censo_2010/grade_estatistica/ge_guia_utilizacao.pdf.

González, L. y Torres, E. (2012) 'Estimaciones de población en áreas menores en América Latina: Revisión de métodos utilizados.', en *Estimaciones y proyecciones de población en América Latina. Desafíos de una agenda pendiente.* Suzana Cavenaghi Rio de Janeiro: ALAP.

Han, P. et al. (2014) Monitoring trends in light pollution in china based on nighttime satellite imagery. *Remote Sensing*. 65541–5558.

Hastie, T. y Tibshirani, R. (1986) Generalized additive models. *Statistical Science*. 1 (3), 297–318.

Hijmans, R. J. (2016) R package version 2.5-8. *Raster: Geographic data analysis and*

*modeling*. [online]. Disponible en: https://CRAN.R-project.org/package=raster.

Howe, A. (2004) *"Assessing the accuracy of australia's small área population estimates, 2001.*

Imhoff, M. et al. (2010) 'Disaggregation of national fossil fuel CO2 emissions using a global power plant database and DMSP nightlight data', en *Proceedings of the 30th Asia-Pacific Advanced Network Meeting*. 2010 Hanoi, Vietnam:.

INE (2005) *Estimaciones y proyecciones de la población de Uruguay (1996-2050) y departamentos (1996-2025). (Revisión 2005).*

INE (2014) *Estimaciones y proyecciones de la población de Uruguay: Metodología y resultados (Revisión 2013).*

INE (1998) *Uruguay: Estimaciones y proyecciones de la población por sexo y edad: Total del país 1950-2050.*

INE (1999) *Uruguay: Estimaciones y proyecciones de la población urbana y rural por sexo y edad 1985-2050.*

James, G. et al. (2013) DOI: 10.1007/978-1-4614-7138-7. *An Introduction to Statistical Learning*. Springer Texts in Statistics. Vol. 103. New York, NY: Springer New York. [online]. Disponible en: http://link.springer.com/10.1007/978-1-4614-7138-7 (Fecha de consulta 20 Diciembre 2017).

Jannuzzi, P. de M. (2005) '*Population projections for small areas: Method and applications for districts and local population projections in brazil*', en 2005 IUSSP. [online]. Disponible en: http://iussp2005.princeton.edu/download. aspx?submissionId=51422.

Judson, D. H. y Swanson, D. A. (2011) 'Estimating Characteristics of the Foreign Born by Legal Status: An Evaluation of Data and Methods', en *Estimating Characteristics of the Foreign-Born by Legal Status*. Dordrecht: Springer Netherlands. págs 1–50. [online]. Disponible en: http://www.springerlink.com/index/10.1007/978-94-007-1272-0_1 (Fecha de consulta 3 Mayo 2017).

Kiran Chad, T. et al. (2009) Spatial characterization of electrical power consumption patterns over India using temporal DMSP-OLS night-time satellite data. *International Journal of Remote Sensing*. 30647–661.

Legates, D. R. y McCabe, G. J. (1999) Evaluating the use of 'goodness-of-fit' Measures in hydrologic and hydroclimatic model validation. *Water Resources Research*. [Online] 35 (1), 233–241. [online]. Disponible en: http://doi.wiley.com/10.1029/

1998WR900018 (Fecha de consulta 31 Mayo 2018).

Legendre, A. M. (1805) *Nouvelles méthodes pour la détermination des orbites des comètes*. F. Didot.

Lesmeister, C. (2017) OCLC: 989783044. *Mastering machine learning with R: Advanced prediction, algorithms, and learning methods with R 3.x*. [online]. Disponible en: https://nls.ldls.org.uk/welcome.html?ark:/81055/vdc_100044650304.0x000001 (Fecha de consulta 30 Enero 2018).

Liaw, A. y Wiener, M. (2002a) Classification and Regression by randomForest. *R News*. 2 (3), 18–22. [online]. Disponible en: http://CRAN.R-project.org/doc/Rnews/.

Liaw, A. y Wiener, M. (2002b) Classification and regression by randomForest. *R News*. 2 (3), 18–22. [online]. Disponible en: http://CRAN.R-project.org/doc/Rnews/.

Liu, Z. et al. (2012) Extracting the dynamics of urban expansion in china using dmsp-ols nighttime light data from 1992 to 2008. *Landscape and Urban Planning*. 106 (1), 62–72.

Long, J. F. (1993) *Post-censal Population Estimates: States, Counties, and Places*. [online]. Disponible en: http://www.census.gov/population/www/documentation/twps0003.html. [online]. Disponible en: http://www.census.gov/population/www/documentation/twps0003.html.

Mennis, J. (2009) Dasymetric Mapping for Estimating Population in Small Areas. *Geography Compass*. [Online] 3 (2), 727–745. [online]. Disponible en: http://doi.wiley.com/10.1111/j.1749-8198.2009.00220.x (Fecha de consulta 12 Diciembre 2017).

Murdock, S. H. y Ellis, D. (1991) *Applied Demography: An Introduction to Basic Concepts, Methods, and Data*. Boulder, Colo: Westview Press Inc.

Nelder, J. y Wedderburn, R. (1972) Generalized linear models. *Journal of the Royal Statistical Society Series A*. (135), 370–384.

Oda, T. et al. (2011) A very high-resolution (1 km × 1 km) global fossil fuel CO2 emission inventory derived using a point source database and satellite observations of nighttime lights. *Atmospheric Chemistry and Physics*. 11543–556. [online]. Disponible en: https://www.atmos-chem-phys.net/11/543/2011/acp-11-543-2011.html.

Pebesma, R. B., E.J. (2005) R News 5 (2). *Classes and methods for spatial data in r*. [online]. Disponible en: http://cran.r-project.org/doc/Rnews/.

Pozzi, F. et al. (2003) Modeling the distribution of human population with nighttime

satellite imagery and gridded population of the world. *Earth Observation Magazine*. 12 (1),.

R Core Team (2016) *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing. [online]. Disponible en: https://www.R-project.org/.

Roger S. Bivand, V. G.-R., Edzer Pebesma (2013) *Applied spatial data analysis with r*. Second edition. Springer, NY. [online]. Disponible en: http://www.asdar-book.org/.

RStudio Team (2012) *RStudio: Integrated development environment for r*. Boston, MA: RStudio, Inc. [online]. Disponible en: http://www.rstudio.com/.

Slocum, T. A. y Slocum, T. A. (eds.) (2009) OCLC: ocn182779739. *Thematic cartography and geovisualization*. Prentice Hall series in geographic information science. 3rd ed. Upper Saddle River, NJ: Pearson Prentice Hall.

Small, C. y Elvidge, C. D. (2013) Night on earth: Mapping decadal changes of anthropogenic night light in asia. *International Journal of Applied Earth Observation and Geoinformation*. 22 (1), 40–52.

Sutton, P. (1997) Modeling population density with night-time satellite imagery and GIS. *Computers, Environment and Urban Systems*. [Online] 21 (3-4), 227–244. [online]. Disponible en: http://linkinghub.elsevier.com/retrieve/pii/S0198971597010053 (Fecha de consulta 2 Mayo 2017).

Sutton, P. C. et al. (2007) *Estimation of gross domestic product at sub-national scales using nighttime satellite imagery*. 85–21.

Sutton, P. C. et al. (2010) 'A 2010 mapping of the constructed surface area density for S.E. Aisa—Preliminary results.', en *Proceedings of the 30th Asia-Pacific Advanced Network Meeting*. 2010 Hanoi, Vietnam:.

Sutton, P. et al. (1997). A comparison of nighttime satellite imagery and population density for the continental us. *Photogrammetric Engineering and Remote Sensing*. 63 (11), 1303–1313.

Texeira Jardim, M. L. (2001) Uso de variables sintomáticas para estimar la distribución espacial de población. aplicación a los municipios de río grande do sul, brasil. *Revista Notas de Población*. 7121–50. [online]. Disponible en: http://www.cepal.cl/publicaciones/xml/3/7223/LCG2114_p2.pdf.

Therneau, T. M. et al. (1997) *An introduction to recursive partitioning using the*

*rpart routines*.

Tin Kam Ho (1998) The random subspace method for constructing decision forests. *IEEE Transactions on Pattern Analysis and Machine Intelligence*. [Online] 20 (8), 832–844. [online]. Disponible en: http://ieeexplore.ieee.org/document/709601/ (Fecha de consulta 22 Marzo 2018).

Wood, S. N. (2017) *Generalized additive models: An introduction with r*. CRC press.

Wood, S. N. (2003) Thin plate regression splines. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*. 65 (1), 95–114.

Wu, J. et al. (2013) Intercalibration of dmsp-ols night-time light data by the invariant region method. *Int. J. Remote Sens.* 347356–7368.

Zhang, H. y Singer, B. H. (2010) *Recursive partitioning and applications*. Springer Science & Business Media.

Zhao, N. et al. (2011) *Net primary production and gross domestic product in China derived from satellite imagery.* 70921–928.