Mapeando el ingreso del Área Metropolitana de Buenos Aires en Alta Resolución: Un enfoque basado en Redes Neuronales aplicadas a Imágenes Satelitales*

Nicolás F. Abbate

Tesis de Maestría

Maestría en Economía

Universidad Nacional de La Plata

Director de tesis: Leonardo Gasparini

Co-director de tesis: Franco Ronchetti

Fecha de defensa: 15/08/2024

Códigos JEL: C45, O18, R12

Mapeando el ingreso del Área Metropolitana de Buenos Aires en Alta Resolución: Un enfoque basado en Redes Neuronales aplicadas a Imágenes Satelitales*

Tesis para optar al Grado de Magíster en Economía, Junio 2024

Nicolás F. Abbate
Universidad Nacional de la Plata
abbatenicolas@gmail.com

Director: Leonardo Gasparini (CEDLAS)
Codirector: Franco Ronchetti (III-LIDI, CICPBA)

***Resumen*—El acceso a datos socioeconómicos desagregados y actualizados es fundamental para la formulación y evaluación de políticas públicas. En este estudio se explora el potencial de las imágenes satelitales de alta resolución en conjunto a técnicas de aprendizaje automático para construir mapas de ingreso con alto grado de desagregación geográfica. Utilizando una red neuronal convolucional (CNN) entrenada con imágenes satelitales del Área Metropolitana de Buenos Aires (Argentina) y datos censales de 2010, se generan estimaciones del ingreso per cápita a nivel de grilla de 50x50 metros para los años 2013, 2018 y 2022, superando la resolución y frecuencia de los datos censales disponibles. El modelo, basado en EfficientnetV2, alcanza un alto nivel de precisión en la predicción del ingreso de los hogares (R^2 del 0.878), superando la capacidad predictiva y mejorando la resolución espacial de otras alternativas utilizadas en la literatura. Este enfoque ofrece nuevas oportunidades para la generación de datos altamente desagregados, facilitando la evaluación de políticas públicas a escala local, generando insumos para una mejor focalización de programas sociales, y reduciendo la brecha de datos en áreas donde estos no se recolectan.**

***Palabras clave*—Indicadores sociodemográficos, Desigualdad, Ingreso per cápita, Imágenes satelitales, Aprendizaje automático**

***Códigos JEL*—C81, C45, R12**

I. Introducción

Los indicadores sociodemográficos son fundamentales para la formulación y evaluación de políticas públicas, ya que permiten medir la eficacia de los programas gubernamentales y el bienestar de la población, mejorando así la transparencia y la rendición de cuentas. Sin embargo, para poder ser utilizados apropiadamente, estos indicadores necesitan estar actualizados y desagregados. Si no se actualizan regularmente, los indicadores no estarán disponibles a tiempo para la toma de decisiones de política pública. Si no tienen un nivel de detalle y desagregación apropiados, no será posible discriminar correctamente los efectos de las diferentes políticas públicas que se implementen.

Dadas las fuentes de datos tradicionales para el análisis socioeconómico —encuestas y censos— obtener simultáneamente datos actualizados regularmente y con una elevada desagregación se vuelve imposible. Por ejemplo, en Argentina, la Encuesta Permanente de Hogares (EPH) se publica trimestralmente y cubre solo los 31 principales aglomerados urbanos, con desagregación limitada al nivel de aglomerado (Instituto Nacional de Estadística y Censos, 2003). Por otro lado, los censos brindan información a nivel de radio censal —que tiene el tamaño de una manzana en áreas de alta densidad poblacional, pero que puede llegar a abarcar varios kilómetros en áreas de baja densidad[1]—, pero solo se realizan cada 10 años. Además, en los censos generalmente no se recopilan datos sobre ingresos o gastos de los hogares, dificultando la estimación directa de indicadores de pobreza monetaria.

Las imágenes satelitales de alta resolución han surgido recientemente como una fuente práctica de información sobre el bienestar, en gran parte gracias a los nuevos avances en algoritmos de visión por computadora. Los avances en el aprendizaje profundo, como las Redes Neuronales Convolucionales (CNN), tienen la capacidad de clasificar imágenes en base a patrones comunes detectados endógenamente, así como también identificar algorítmicamente objetos como automóviles, área de construcción, carreteras, cultivos y tipo de techo. En ambos casos, la identificación de estas características u objetos es importante ya que están fuertemente correlacionados con la riqueza y los ingresos locales.

A pesar de estos avances, las estimaciones de bienestar generalmente se limitan a áreas agregadas, como ciudades o provincias. Esto se contrapone con lo que disciplinas como la climatología o la demografía han generado recientemente con

*El presente trabajo es una derivación del documento de trabajo de Abbate, Gasparini, Gluzmann, Montes-Rojas, Sznaider y Yatche (2023), "Ingreso Estructural Por Area Geográfica: una aplicación para Argentina", presentado en la LVIII Reunión Anual de la Asociación Argentina de Economía Política. Agradezco especialmente a mis directores Leonardo y Franco, quienes lograron que este trabajo fuera posible. Agradezco los comentarios recibidos en el seminario de avance de tesis. Agradezco a la Comisión Nacional de Actividades Espaciales (CONAE) por facilitar las imágenes satelitales utilizadas en este trabajo. Agradezco a mi familia y amigos, por acompañarme e inspirarme continuamente. Los errores y omisiones son de mi exclusiva responsabilidad

[1]Los radios censales son una unidad geográfica que agrupa en promedio a 300 viviendas en las ciudades y abarcan la totalidad del territorio nacional, lo que se traduce en una muy importante variabilidad en su tamaño.

insumos similares: la generación de estimaciones en forma de grilla. Los datos en formato de grilla permiten democratizar su acceso para una amplia diversidad de aplicaciones que muchas veces exceden el objetivo original de los generadores de estas bases. Por ejemplo, el proyecto Gridded Population of the World (Center for International Earth Science Information Network (CIESIN), 2018), estima la densidad poblacional para todo el mundo sobre una grilla de aproximadamente 1km de lado; al año 2024, esta base de datos se encuentra citada por más de 8,000 artículos revisados por pares.

El objetivo de este trabajo es, utilizando una red neuronal convolucional aplicada a imágenes satelitales de alta resolución (0.5 metros por pixel, similar a los datos de Google Maps) del Área Metropolitana de Buenos Aires en 2013, generar estimaciones del ingreso per cápita espacial de los hogares con una muy alta desagregación: una grilla con celdas de aproximadamente 50x50 metros. Las estimaciones se realizan para el año 2013, 2018 y 2022. Los mapas generados con esta metodología se encuentran disponibles al público para su uso en este repositorio[2]. Estos mapas presentan una mejora sustancial sobre la resolución de los datos originales. Los datos utilizados provienen del Censo 2010 y de la Encuesta Permanente de Hogares del segundo semestre de 2010, que aportan la información del ingreso per cápita geográfico para todo el AMBA, en conjunto con la información no estructurada disponible en las imágenes satelitales de muy alta resolución.

Con el propósito de construir estos mapas, se entrenó una red neuronal convolucional de la familia de *EfficientNetV2* (Tan and Le, 2019). Esta clase de modelos presentan la marcada ventaja de que permiten identificar, de forma endógena, las características explicativas de las imágenes que generan un mejor desempeño para la predicción; es decir, no se imponen supuestos *a priori* sobre las características que determinan el ingreso en las imágenes satelitales.

El modelo resultante permite generar predicciones con una precisión muy elevada a nivel de radio censal, alcanzando un R^2 del 0.878 sobre el conjunto de prueba. Se evaluó la consistencia de los mapas construidos comparando los resultados con información censal, con información de la Encuesta Permamente de Hogares, y evaluando casos testigo en áreas que presentan discontinuidades marcadas del ingreso, como asentamientos informales. En todos los casos, el desempeño del modelo es muy positivo, superando los resultados obtenidos mediante metodologías alternativas presentadas en la literatura.

La construcción de un modelo que prediga el ingreso medio de los hogares a partir de únicamente una imagen satelital del área de interés implica una serie de aplicaciones sumamente relevantes. En caso de desarrollarse un modelo generalizado para muchas ciudades diferentes —y no limitado únicamente en el AMBA—, podría generarse nueva información para áreas urbanas donde no hay información disponible, mediante la predicción del modelo sobre las imágenes satelitales de dichas áreas. Esto podría ayudar a reducir la brecha de datos existente en zonas rurales y de bajos ingresos, a donde no alcanzan muchas de las encuestas (Burke et al., 2021).

[2]https://doi.org/10.5281/zenodo.11200070

En segundo lugar, los mapas generados con esta técnica podrían facilitar el desarrollo de evaluaciones de impacto localizadas. Por ejemplo, resultaría significativamente más simple evaluar el impacto de nuevas redes cloacales o de agua corriente en el desarrollo urbano de una zona determinada, utilizando únicamente las predicciones del modelo antes y después del tratamiento, si se selecciona un grupo de control apropiado.

Finalmente, la mejoría de la resolución espacial y temporal de los mapas de ingreso facilitaría la tarea de evaluación de la focalización de programas sociales, ya que en muchos casos se dispone de información geolocalizada del lugar de residencia de los beneficiarios. Utilizando la ubicación de cada beneficiario, es posible imputar un ingreso "geográfico" basado en su lugar de residencia. Entonces, estas evaluaciones ya no dependerían únicamente de datos administrativos potencialmente desactualizados o de estimaciones *ad-hoc* basadas en encuestas. Además de esto, estas técnicas podrían ayudar a desarrollar mejores estrategias de focalización de programas sociales, permitiendo nuevos criterios de exclusión que podrían aplicarse dentro de los *proxy mean tests* (Grosh and Baker, 1995), o como criterios independientes de exclusión (Smythe and Blumenstock, 2022).

Este trabajo se relaciona directamente con la literatura que busca predecir indicadores sociales a nivel espacial utilizando "proxies" del indicador, fuentes de datos alternativas que presenten una fuerte correlación con el indicador a predecir. Entre muchas aplicaciones, se destacan el uso de "luces nocturnas"satelitales para predecir el crecimiento económico (Henderson et al., 2012) y la desigualdad (Ciaschi, 2021), datos de teléfonos móviles para predecir la pobreza y la riqueza (Blumenstock et al., 2015; Steele et al., 2017) o información extraída de redes sociales para predecir el desarrollo económico (Sheehan et al., 2019; Weber et al., 2018).

Dentro de esta literatura, recientemente ha cobrado mucha relevancia aquella que utiliza imágenes satelitales diurnas de alta resolución para la predicción de indicadores sociales como la pobreza (Engstrom et al., 2022; Jean et al., 2016), el ingreso (Khachiyan et al., 2022; Piaggesi et al., 2019; Rolf et al., 2021) o la riqueza (Chi et al., 2022; Yeh et al., 2020). Si bien el desempeño de este tipo de modelos suele ser bueno, generalmente el área sobre la cual se generan las predicciones de los indicadores suele ser muy amplia, típicamente estimando los indicadores para ciudades enteras; en los casos de mayor desagregación, predicen el indicador económico en celdas de 1km de lado (ver, por ejemplo, (Piaggesi et al., 2019)). La aplicación de la metodología que se propone, que combina información censal con imágenes de alta resolución utilizando una arquitectura de frontera, permite alcanzar un modelo superador tanto en su resolución espacial (mayor desagregación del indicador) como en su capacidad predictiva (menor error).

Este documento se estructura de la siguiente forma. En la sección II se presenta la metodología general utilizada en este trabajo. Las secciones siguientes desarrollan en detalle la metodología utilizada, especificando la construcción de la base de datos (sección III), la implementación y entrenamiento de la red neuronal convolucional (sección IV) y el procedimiento

(a) Barrio Los Ombúes, Tigre. (b) Barrio San Justo, La Matanza. (c) Barrio 31, CABA.

Figura 1: Ejemplos de imágenes satelitales de 200x200 metros.

para construir los mapas del ingreso para 2013, 2018 y 2022 (sección V). En la sección VI se presenta los principales resultados del trabajo. En la sección VII se discuten la interpretación del modelo, así como también algunas limitaciones y extensiones del mismo. La sección VIII presenta las principales conclusiones del trabajo.

II. Diseño Metodológico

El presente trabajo busca aportar al campo emergente de la estimación de mapas de bienestar utilizando imágenes satelitales. Para ello, se busca construir una serie de mapas que estiman el ingreso per cápita de los hogares a nivel submunicipal, en celdas de aproximadamente 50x50 metros. Las estimaciones se realizan para el año 2013, 2018 y 2022. Para predecir el ingreso per cápita con este nivel de desagregación, se utiliza una red neuronal convolucional, que permite extraer endógenamente características observables de las imágenes satelitales que predigan consistentemente el ingreso per cápita promedio de los hogares que viven en esa área.

Los algoritmos de inteligencia artificial, y particularmente los modelos de aprendizaje profundo, han permitido resolver problemas complejos que décadas de investigación no habían podido resolver; entre ellos, el reconocimiento y clasificación de imágenes mediante estos algoritmos, han generado una revolución tecnológica (LeCun et al., 2015). Antes de las Redes Convolucionales se utilizaban métodos manuales para la extracción de características de imágenes, lo que implicaba un importante esfuerzo e inversión de recursos para lograrlo. En contraposición a esto, las redes neuronales permiten extraer aquellos patrones presentes en las imágenes que presenten una mayor correlación con el indicador de interés.

La principal motivación para entrenar un modelo de inteligencia artifical que permita predecir el ingreso geográfico a partir de una imagen satelital de la zona de interés surge de que, en muchos casos, es posible reconocer a simple vista las regiones de ingresos altos, medios y bajos, sin ningún tipo de información añadida (Figura 1). En este sentido, el objetivo es encontrar un modelo que pueda representar esa función o mapeo entre características observables de las imágenes —como la forma de las construcciones, los materiales, el asfalto en las calles, la presencia de espacios verdes— y el ingreso medio de las personas que habitan esa zona, sin definir ninguna forma funcional, especificación o conjunto de variables relevantes a priori. La red neuronal, una vez entrenada, podrá identificar endógenamente cuáles de estas características son las que mejor predicen el ingreso.

Para lograr la extracción de características de las imágenes, se "entrena" al modelo utilizando una serie de ejemplos, a partir de los cuáles encontrará el conjunto de parámetros que genere mejores predicciones. El conjunto de ejemplos, que de aquí en adelante se mencionará como *conjunto de entrenamiento*, está compuestos por (i) un conjunto de imágenes, y (ii) una variable que asigne a cada imagen el indicador que se busca predecir. En este caso, las imágenes utilizadas son las capturas satelitales del Área Metropolitana de Buenos Aires. La variable a predecir es el ingreso per cápita promedio de los hogares que viven en el área de donde fue extraída la imágen. Una vez entrenado el modelo, este puede generar predicciones en cualquier conjunto de imágenes, por lo que puede ser utilizado para predecir el ingreso en otras fechas —y por lo tanto generar series de tiempo del ingreso geográfico— y de otras ciudades con características similares.

La estrategia de estimación de este trabajo puede descomponerse en tres instancias. En la Figura 2 se representa esta estrategia de estimación, desde las bases de datos utilizadas a la sección de resultados. Cada una de estas etapas se desarrollan de forma particular en las secciones que siguen a continuación. El paso (1), desarrollado en la sección III, consiste en la construcción de una base de datos apropiada para el entrenamiento de una red neuronal. Por un lado, esto implica estimar el ingreso per cápita medio por radio censal, ya que esta información no está disponible de forma desagregada para la Argentina. La estimación se realiza con una técnica estándar de estimación por área pequeña (*small area estimation*), utilizando una variación de la metodología de Elbers et al. (2003) y Gasparini et al. (2022) para combinar los datos de la Encuesta Permanente de Hogares del segundo semestre 2010 y con los datos del Censo Nacional de Población, Hogares y Viviendas 2010. Por el otro, se construye una base de imágenes satelitales para el AMBA del

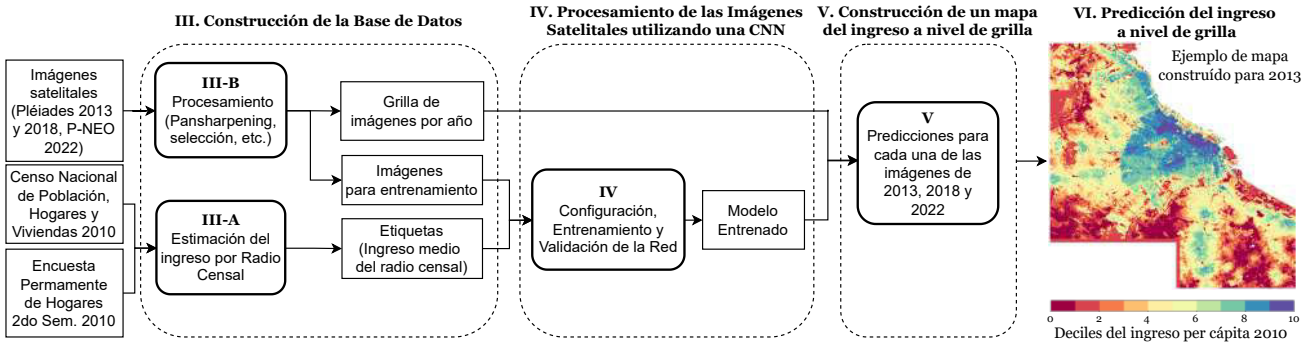


Figura 2: Estrategia de estimación, por secciones de este trabajo.

año 2013 con imágenes cuadradas, el formato requerido por la red neuronal para su procesamiento. El paso (2), desarrollado en la sección IV es el entrenamiento de la red neuronal convolucional (CNN, por sus siglas en inglés) que permita relacionar estas imágenes satelitales con el ingreso estimado para 2010. Finalmente, el paso (3), desarrollado en la sección V, consiste en generar una serie de estimaciones del ingreso en forma de grilla de $50x50mts$ del AMBA para los años 2013, 2018 y 2022, utilizando el modelo entrenado previamente. En otras palabras, se genera una serie de mapas de muy alto nivel de desagregación utilizando las predicciones del modelo para las imágenes de 2013, 2018 y 2022.[3]

III. Construcción de la Base de Datos

III-A. Construcción del indicador: Una estimación del Ingreso por Radio Censal por área pequeña

Los datos del ingreso per cápita de cada radio censal se estiman utilizando los microdatos del Censo Nacional de Población, Hogares y Viviendas 2010, en conjunto con los de la Encuesta Permanente de Hogares (EPH) del segundo semestre de 2010, siguiendo una adaptación de la metodología propuesta por Elbers et al. (2003). Los datos de la EPH se procesan siguiendo la metodología utilizada por Tornarolli (2018). A pesar de que los datos del censo contienen información detallada sobre prácticamente la totalidad de los habitantes del país, en el cuestionario no se incluye información sobre el ingreso de los hogares. El método utilizado permite combinar información censal espacialmente desagregada con información de encuestas de hogares, que tienen un tamaño muestral reducido, pero que contienen información de los ingresos de los hogares.

En el trabajo original de (Elbers et al., 2003), los autores plantean una *estimación de área pequeña* que consiste en: (i) seleccionar un indicador de interés que se encuentre disponible en una encuesta pero no en el censo, (ii) identificar covariables que se relacionen con el indicador y que se encuentren tanto en el censo como en la encuesta, (iii) estimar un modelo lineal de mínimos cuadrados generalizados sobre la encuesta, que relacione el indicador de interés con las covariables seleccionadas, y (iv) generar predicciones del indicador en el censo, utilizando las covariables censales disponibles y las estimaciones de los parámetros del modelo del punto previo. Estas predicciones tienen la ventaja de estar geográficamente localizadas, ya que típicamente los censos permiten una desagregación espacial mucho mayor que las encuestas de hogares. En general, luego de esta predicción se agregan los indicadores al mínimo nivel espacial disponible, lo que permite construir mapas con las estimaciones correspondientes.

Para realizar la estimación espacial del ingreso, entonces, se estimó un modelo lineal sobre los microdatos de la Encuesta Permanente de Hogares del 2do semestre de 2010, estimando la relación entre el logaritmo del ingreso per cápita familiar de cada hogar i (y_i^{eph}) y un vector de variables observables (X_i^{eph}). Las variables observables utilizadas para la predicción incluyen el género, educación y edad del jefe o jefa del hogar, binarias que identifican la calidad de la vivienda (materiales, acceso al agua, clocas, baño, si es propietario, si tiene materiales precarios, si es vivienda precaria) y la cantidad de miembros del hogar. Por lo tanto, el modelo correspondiente es de la forma:

$$y_i^{eph} = \beta^{eph} X_i^{eph} + \varepsilon_i \tag{1}$$

Una vez estimado el modelo, se guardan las estimaciones del vector de parámetros β y se replica el modelo con las covariables del censo X_{ij}^c. En otras palabras, se utiliza la relación estimada en la EPH entre las variables observables de los hogares y el logaritmo del ingreso para predecir el dato faltante en el censo. La única diferencia es que, en el caso del censo, el hogar i además tiene una identificación con un radio censal j, que será posteriormente utilizado para asignar una dimensión geográfica a la estimación. Específicamente, se computa la siguiente relación:

$$\hat{y}_{ij} = \beta^{eph} X_{ij}^c \tag{2}$$

Finalmente, para construir el indicador del ingreso por radio censal (en logaritmos) se calcula la media de la variable predicha de cada uno de los radios censales. Por lo tanto, definiendo n_j como el número de hogares contenidos en el

[3]El uso de imágenes satelitales de los años 2013, 2018 y 2022 se debe únicamente a su disponibilidad. Las imágenes utilizadas fueron adquiridas por la CONAE para esos años. Si bien existen alternativas de uso libre, en todos los casos la resolución es muy inferior a la utilizada. A modo de ejemplo, el satélite Sentinel 2 provee de forma gratuita imágenes con un máximo de resolución de 10 metros por pixel. Comparado con las imágenes de 0.5 metros por pixel utilizadas, la resolución es 20 veces menor.

radio censal j, el *ingreso estimado* del radio censal j ($\hat{Y}_j$) queda definido como:

$$\hat{Y}_j = \sum_{i \in j} \frac{\hat{y}_{ij}}{n_j} \tag{3}$$

Con esta construcción, es posible generar mapas del ingreso estimado por radio censal; es decir, una estimación por área pequeña. La Figura 3 muestra la distribución espacial del logaritmo del ingreso per cápita promedio por radio censal en el AMBA, en deciles. Este ingreso estimado es el insumo principal, junto a las imágenes satelitales, que se utilizará para el entrenamiento del modelo de visión por computadora.

A diferencia del método planteado por Elbers et al. (2003), en vez de realizar una estimación por mínimos cuadrados generalizados, se realiza una estimación por mínimos cuadrados ordinarios. Esta diferencia se debe a que los autores cuentan una encuesta que identifica a los hogares con la mínima unidad de agregación del censo. En cambio, utilizando la EPH, no se disponibiliza en los datos de los hogares a qué radio censal pertenecen, por lo que no es posible estimar la matriz de varianzas y covarianzas necesarias para una estimación de GLS.

A pesar de estas dificultades respecto a la calidad de la información original, se espera que el resultado de estas estimaciones presente un sesgo bajo. Esto se debe a que, si bien las estimaciones de cada hogar estarán sesgadas, es esperable que la agregación del ingreso por radio censal mitigue el sesgo, ya que los errores no tienen una correlación perfecta y en buena medida se contrarestarían. Es esperable que el sesgo sea muy bajo en buena parte de la distribución, donde las características observables explican el ingreso relativamente bien, mientras que en los percentiles superiores, es posible que exista un sesgo singificativamente mayor.[4] Por otra parte, Gasparini et al. (2022) utilizan una implementación similar para estimaciones de pobreza, y argumentan que utilizar estimaciones de este tipo a lo largo del tiempo pueden permitir estimar más apropiadamente a la población vulnerable.

Más allá de esta discusión, es importante destacar que la aplicación de modelos de visión por computadora para la estimación del ingreso geográfico es independiente de la validez de la estimación de área pequeña propuesta, ya que una estimación geográfica del ingreso mejorada utilizando otra metodología solo implicaría que los resultados finales del algoritmo de inteligencia artificial mejoren. Por ejemplo, el método planteado podría aplicarse a otras variables con menor error de medición, como el precio por metro cuadrado promedio de las viviendas. Asimismo, en países donde existan datos geográficos desagregados del ingreso, esa variable podría usarse de forma directa en las imágenes, sin la necesidad de estimar el ingreso con el método planteado en esta sección.

[4]En este sentido, es esperable que algunos radios censales de altos ingresos —como los que se encuentran dentro de barrios cerrados de alto poder adquisitivo— el ingreso estimado por área pequeña no los identifique como parte de los últimos percentiles. En el mapa generado, esto se evidencia en la zona de Nordelta, donde muchos radios censales de altos ingresos tienen un ingreso estimado perteneciente al decil 9 u 8, cuando, por los costos que implica una vivienda allí, uno esperaría que pertenezcan al decil 10.

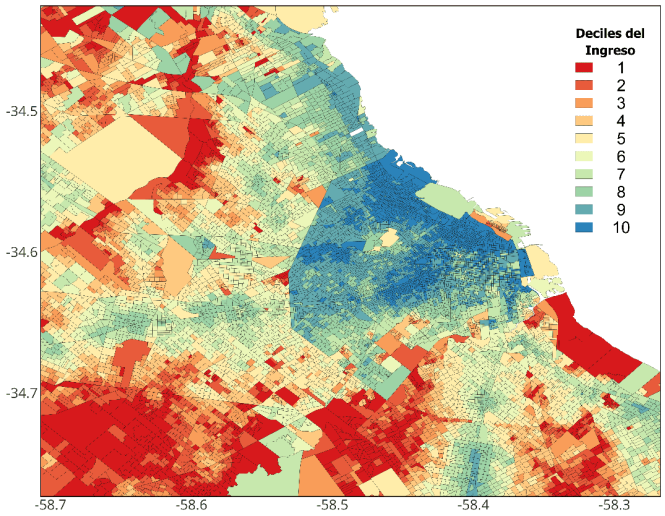


Figura 3: Distribución espacial del ingreso del AMBA, datos a nivel de radio censal.

III-B. Construcción del conjunto de imágenes: de radios censales a imágenes

Las imágenes satelitales utilizadas son imágenes diurnas de reflectancia de la superficie terrestre generadas por la constelación de satélites Pléiades del *Centre national d'études spatiales*. La constelación Pléiades tiene la característica de generar imágenes de aproximadamente 50 centímetros por pixel con una frecuencia de revisita de 26 días. Si bien las imágenes no son de uso público, el acceso a estas fue provisto por la Comisión Nacional de Actividades Espaciales (Argentina). En comparación, las imágenes satelitales de libre disponibilidad de mayor resolución disponibles son las del satélite Landsat 8, que tienen una resolución máxima 30 veces menor (15 metros por pixel).

Se utilizó una composición de imágenes de parte del Área Metropolitana de Buenos Aires, capturadas el 5 y 7 de febrero de 2013. La fecha de las imágenes es la más cercana al censo realizado el 27 de octubre de 2010. Asimismo, para aportar variedad al entrenamiento del modelo, también se incoporaron al entrenamiento las imágenes de 2018 y 2022 (en el Anexo A se muestra empíricamente cómo incluir estas imágenes mejoran el desempeño del modelo). Las imágenes del 2018 son del mismo satélite, mientras que las del 2022 son de su versión más reciente, Pléiades Neo, con características muy similares. En la Figura 6a se puede observar el área de cobertura de las imágenes utilizadas, las cuales abarcan parcialmente la totalidad del AMBA.

Las imágenes están compuestas por cuatro bandas espectrales, correspondientes al color azul (430 - 550nm), verde (500 - 620nm), rojo (590 - 710nm) e infrarrojo cercano o NIR, por sus siglas en inglés (740 - 940nm). Utilizando estas imágenes multiespectrales, generalmente se construyen imágenes a color, combinando las capas roja, azul y verde, así como también pueden derivarse indicadores de vegetación y estrés vegetal de la banda infrarroja cercana, como el NDVI (Normalized Difference Vegetation Index) (Huang et al., 2021). Para la red neuronal se utilizan las cuatro bandas de

forma conjunta, sin proponer una relación o transformación a priori entre las bandas. Además, está disponible la banda pancromática (470 - 830nm), que genera una imagen de mayor resolución utilizando la totalidad del espectro visible, a costa de tener una escala monocromática.

A las imágenes originales se les equalizó el histograma de colores y se les aplicó la transformación de Brovey (Vrabel, 1996) para mejorar la resolución de las imágenes de 1 metro por pixel (resolución original de las bandas roja, verde, azul y NIR, es decir, de las bandas multiespectrales) a 0.5 metros por pixel (resolución de la banda pancronmática). Este algoritmo combina la banda pancromática con las bandas espectrales de menor resolución para aumentar la resolución de las bandas espectrales a la resolución de la banda pancromática. La Figura 4 muestra un ejemplo de este procesamiento, generalmente denominado *pansharpening*, que permite que las imágenes resultantes sean más nítidas y, por lo tanto, más fácilmente detectables para el modelo de visión por computadora.

Una de las dificultades más importantes a la hora de implementar un modelo de procesamiento de imágenes como el estipulado utilizando información del censo es la dimensión espacial sobre la cuál trabajar. Como se aprecia en la Figura 3, los radios censales no tienen un tamaño constante, sino que, por el contrario, se seleccionan buscando que los radios tengan un número similar de hogares. Entonces, a medida que la densidad poblacional es menor, el tamaño de los radios censales crece. En este sentido, es necesario realizar un preprocesamiento del conjunto de datos para normalizar la información geográfica a un tamaño homogéneo.

El tamaño de las imágenes no es una decisión obvia *a priori*. Por un lado, utilizar imágenes con un tamaño mucho menor al de los radios censales (el nivel al que se encuentra la información del ingreso) puede implicar una significativa mejora en la resolución de los mapas generados, a costa de una posible perdida en la calidad de esas estimaciones. De hecho, si se utilizan tamaños de imágenes muy pequeñas (por ejemplo, predecir el ingreso sobre imágenes de 10×10 metros) es probable que el modelo entrenado no logre extraer características relevantes que se correlacionen con el ingreso, debido a que muchas de estas exceden el tamaño de la imagen. Sin embargo, es posible que el modelo permita desagregar los indicadores, ya que dentro de los radios censales las características observables de las imágenes son muy uniformes en la mayoría de los casos. Siguiendo esta lógica, es posible suponer que, cuando existan discontinuidades en estas características, como un asentamiento informal al costado de una vía de tren, el modelo podría aprender a detectarlas, reconociendo el ingreso promedio de radios censales de asentamientos populares con similares características.

En este trabajo se utilizaron imágenes de 50×50 metros, lo que implica cuadruplicar la resolución de la estimación del ingreso para los radios censales de las zonas de mayor densidad poblacional. En píxeles, las imágenes generadas tienen una resolución de $128 \times 128px$. A estas imágenes se les agrega, para agregar contexto a la imagen, una captura de 200×200 metros reescalada a $128 \times 128px$ y centrada en el mismo punto que la imagen de 50×50. Como resultado, las imágenes tendrán $128 \times 128px$ y 8 bandas, donde las bandas 1 a 4 de la imagen representan las capas roja, verde, azul y NIR de la imagen de 50×50 metros, y las bandas 5 a 8 representan las capas roja, verde, azul y NIR de la imagen de 200×200 metros. La decisión del tamaño de imágenes se realizó comparando la precisión de diferentes modelos entrenados con distintos tamaños de imágenes. Esta selección, y la construcción de las imágenes combinadas de diferentes tamaños, se detallan en el Anexo B. Entre las opciones se consideraron utilizar imágenes de 50×50, 100×100 y 200×200 metros, incluyendo además combinaciones de diferentes tamaños. En la selección de tamaños de imágenes, se encontró que al utilizar las imágenes combinadas de 50×50 metros y 200×200 metros se obtuvo un mejor desempeño del modelo, y por ello tales imágenes son las utilizadas a lo largo de este trabajo.

Durante el entrenamiento del modelo, el conjunto de datos utilizado se construye de forma dinámica, seleccionando para cada radio censal una muestra aleatoria de 5 imágenes en cada iteración (época). Para construir cada imagen durante el entrenamiento se selecciona, primero, un punto al azar dentro del radio censal. Luego, la imagen se construye generando un cuadrado de $50x50$ (y $200x200$) metros, teniendo tal punto como centro. Este proceso se repite 5 veces por cada radio censal, en cada época. Debido al tamaño de las imágenes y la aleatoriedad del centro de la imágen, muchas de estas contendrán información parcial de sus radios contiguos, lo que podría aportar información relevante en la estimación, sobre todo cuando existen saltos discretos en el ingreso, o características relevantes como autopistas o ríos. La selección de un número constante de imágenes por cada radio censal, en vez de la construcción de una grilla del area del AMBA, busca evitar que en el entrenamiento queden sobrerrepresentadas imágenes de baja densidad poblacional. A su vez, como se aleatorizan las imágenes en cada época, el modelo se ve enfrentado constantemente a imágenes diferentes de los radios censales. Se aplicaron además otras técnicas estándar de *data augmentation* modificando aleatoriamente el brillo, contraste y orientación de las imágenes (Shorten and Khoshgoftaar, 2019).

A cada imagen se asigna el ingreso per cápita promedio del radio censal a la cual pertenece, estimado por área pequeña según la metodología planteada en la sección III-A. Esta decisión está motivada en la elevada correlación espacial observada en el mapa del ingreso per cápita como en la homogeneidad de las imágenes de barrios o zonas con ingresos similares. Si bien existen variaciones relevantes dentro de algunos radios censales, en general estos tienden a ser visualmente homogéneos, por lo que el valor del ingreso promedio es generalmente representativo de cualquier subconjunto del radio censal.

La Figura 5 muestra un ejemplo de dos radios censales sobre los cuales se seleccionaron aleatoriamente 4 imágenes de cada una. A la izquierda, se muestran los radios censales (rojo) con dos niveles de ingreso diferentes. A la derecha se muestran las imágenes con las que se alimenta el modelo, y se muestra el ingreso de cada una de las imágenes en su esquina superior izquierda.

Figura 4: Mejoramiento de la resolución de las imágenes satelitales (*Pansharpening*).

(a) Multiespectral (MS). (b) Pancromática. (c) MS Mejorada (Pansharpen).

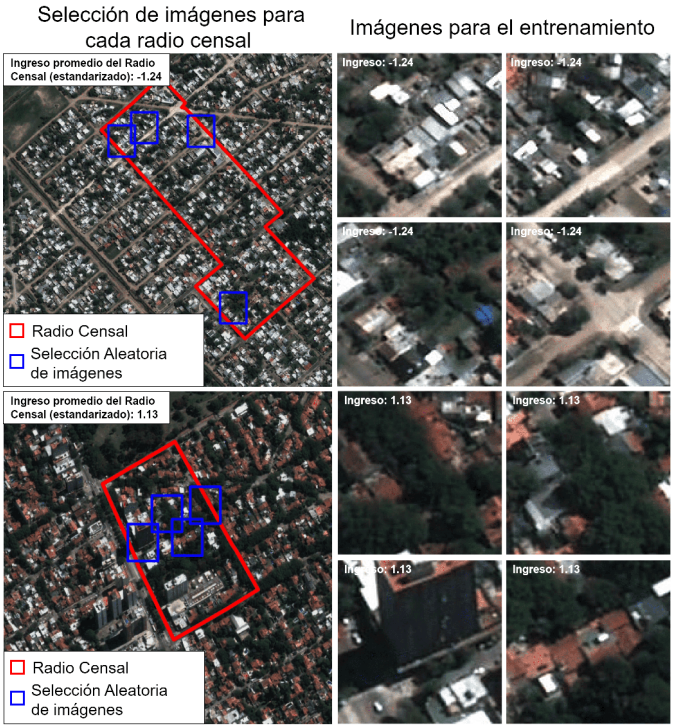


Figura 5: Construcción del conjunto de imágenes de 128x128px (50x50mts) a partir de los datos de un radio censal.

Fuente: Imagenes satelitales provistas por la CONAE.

IV. Procesamiento de las Imágenes Satelitales utilizando una Red Neuronal Convolucional

Las Redes Neuronales Convolucionales (CNN, por su sigla en inglés), así como también las Redes Neuronales en general, tienen la relevante característica de ser un aproximador universal de funciones; es decir, que dada una red lo suficientemente grande, es posible aproximar la función que mapea el conjunto de imágenes con la variable a predecir, cualquiera sea su forma (Yarotsky, 2018). En este sentido, en este trabajo se asume la existencia de una función que relaciona —al menos de forma imperfecta— el ingreso per cápita de los hogares con características observables en las imágenes satelitales.

Para aproximar esta función que relaciona las imágenes con el ingreso, las redes neuronales detectan características y patrones relevantes en los datos de entrada sin la necesidad de una intervención humana, lo que las hace ideales para tareas de procesamiento de imágenes complejas. A diferencia de algunos modelos utilizados en la literatura (por ejemplo, Jean et al. (2016) y Engstrom et al. (2022)), las CNN no requieren una modelización específica de la relación entre el indicador y las características, ni supone a priori cuáles de las características observadas son las más relevantes. Además, pueden capturar relaciones no lineales. Como contrapartida, este tipo de modelos requieren un mayor trabajo para su interpretación ya que su proceso de aprendizaje se basa en la identificación de patrones a través de múltiples capas que se conectan mediante una gran cantidad de parámetros.[5]

El objetivo es que la red extraiga información económica que está latente en las imágenes satelitales. La distribución de características observables mediante imágenes satelitales varía enormemente dentro de una misma ciudad: más vegetación y presencia de tierra en áreas suburbanas; más asfalto y cemento alrededor de las autopistas; edificaciones con predominancia de chapa y calles de tierra en asentamientos informales. Las formas de estas características exhiben una variación igualmente amplia: casas con amplios jardines y espacios verdes en zonas suburbanas de ingresos medios-altos; rejillas compactas e interconectadas en centros urbanos (Pesaresi et al., 2016; Ural et al., 2011). Es esta complejidad la que hace poderosa a una red neural: la red aprende a relacionar la compleja asignación de materiales y formas con el nivel de ingreso

[5]Este problema es altamente relevante para tareas donde la variable predicha es inobservable, por ejemplo, la probabilidad de repago de un crédito. Sin embargo, para el problema de este trabajo, esta dificultad es menos importante porque es más fácil contrastar las predicciones del modelo con el ingreso de los hogares, al menos de forma aproximada, basándose en otros indicadores económicos alternativos.

(Khachiyan et al., 2022).

Las redes neuronales consisten en una secuencia de capas en las que se implementa una transformación no lineal parametrizada de sus entradas —las imágenes— para generar una predicción —el ingreso per cápita de los hogares que viven en esa zona. La cantidad de parámetros del modelo puede ser muy alta; por ejemplo, la arquitectura que se utiliza en este trabajo, denominada *EfficientNetV2S*, tiene 21.6 millones de parámetros (Tan and Le, 2019).[6] Los parámetros generalmente se inicializan de forma aleatoria, generando, lógicamente, predicciones muy malas. A lo largo del entrenamiento, los parámetros se van ajustando en pos de mejorar la precisión de las predicciones realizadas sobre un conjunto de entrenamiento. En la literatura de aprendizaje automático, esta forma de entrenar los modelos se denomina *aprendizaje supervisado* (Prince, 2024). El aprendizaje supervisado es actualmente el paradigma más difundido en inteligencia artificial y funciona como base de la gran mayoría de los modelos de procesamiento de imágenes.

Por lo tanto, la extracción de características de las imágenes se realiza mediante un algoritmo de aprendizaje supervisado, que "entrena" a la red neuronal. En el caso de procesamiento de imágenes, el aprendizaje supervisado utiliza como insumos principales (i) un conjunto de imágenes —idealmente grande—, y (ii) una variable que asigne a cada imagen el indicador que se busca predecir. En este trabajo se utilizan imágenes satelitales del AMBA, junto al ingreso per cápita promedio de los hogares que viven en el área de esas imágenes, generadas según el método de estimación descripto en la sección anterior.

El aprendizaje supervisado consiste en utilizar las bases de datos mencionadas para ajustar los parámetros del modelo de forma tal que se minimice el error en las predicciones. Durante el entrenamiento, iterativamente se muestra al algoritmo las imágenes del conjunto de entrenamiento y éste produce una predicción del indicador para cada una de ellas. El objetivo es que el valor de la predicción sea lo más parecido al valor de la variable explicada. Para medir esta diferencia entre el valor predicho y el real, se computa una función de pérdida —o función objetivo— que mida estos errores. El algoritmo, entonces, buscará minimizar la función de pérdida ajustando sus parámetros iterativamente. Para lograr un ajuste adecuado del vector de parámetros, el algoritmo calcula un vector de gradiente que, para cada parámetro, indica en qué medida hay que ajustarlo marginalmente para reducir el error. Luego, se procede a ajustar el vector de parámetros en la dirección opuesta al vector de gradiente. En la práctica, la forma típica de resolver este tipo de problemas implica ir ajustando los parámetros iterativamente, computando el gradiente medio para una muestra reducida de ejemplos en cada iteración. Cada iteración completa sobre el conjunto de datos se denomina "época".[7] Un modelo puede requerir un número importante de épocas para su entrenamiento, pudiendo llegar a los cientos en algunos casos. Finalmente, una vez encontrado un vector óptimo de parámetros, se evalúa la capacidad de generalización del algoritmo frente a un conjunto de nuevo, típicamente denominado *conjunto de prueba*.

Dentro de los algoritmos de inteligencia artificial se destacan, principalmente por su uso generalizado para procesamiento de imágenes, las redes neuronales convolucionales. En estos algoritmos, las entradas de la primera capa son las imágenes en forma matricial. La salida de la primera capa se utiliza como entrada para la segunda capa y así sucesivamente. La transformación implementada por cada capa es típicamente una operación de convolución o agrupamiento (Goodfellow et al., 2016), aunque más recientemente modelos como el que se utilizan aquí presentan transformaciones más complejas (Tan and Le, 2019). La salida de cada capa es otra "imagen", de menor resolución que la original, en la que los píxeles representan un resumen de la información de la capa previa; es decir, en cada capa la red extrae las características principales de la imagen anterior. Progresivamente, el modelo va reduciendo el tamaño de la imagen y disminuye la resolución del mapa de características para finalmente relacionarlos con la variable que se busca predecir.

Es importante destacar que las CNN son robustas a datos "ruidosos" en el entrenamiento; es decir, que si los errores en la variable a predecir son aleatorios, las CNN tienen la capacidad de generalizar apropiadamente la función latente (Rolnick et al., 2017). Esta propiedad es relevante porque, como se mencionó en la sección III-B, la variable utilizada no se corresponderá inequívocamente a cada una de las imágenes, sino que en en todos los casos tendremos imágenes con un ingreso repetido. Esto se debe a que se seleccionarán 5 imágenes diferentes de cada radio censal —algunas de partes más ricas del radio, otras más pobres— todas con el mismo valor de variable: el ingreso medio del radio censal. La capacidad de poder generalizar el modelo en presencia de estas dificultades resulta clave. Sin embargo, debe destacarse que la capacidad de aprendizaje del algoritmo sí depende de la validez de los datos de entrada: si existen sesgos sistemáticos en la estimación del ingreso per cápita por radio censal, entonces el modelo replicará esos sesgos en sus resultados finales. Por lo tanto, es sumamente relevante desarrollar técnicas que permitan mejorar la precisión y validez de las estimaciones de área pequeña.

[6] En el ámbito del *Machine Learning*, el término "arquitectura" se refiere a la estructura general de un modelo de aprendizaje automático. Define la disposición de los componentes del modelo, la forma en que interactúan y el flujo de datos entre ellos. Por ejemplo, una arquitectura muy conocida es la arquitectura de red neuronal convolucional presentada en el trabajo seminal de LeCun et al. (1998) que logra reconocer y clasificar números escritos a mano. Esta arquitectura está compuesta por siete capas, donde la primera, la tercera y la quinta capa realizan una convolución sobre la imagen con un tamaño de 5x5 píxeles, la segunda y la cuarta son capas de agrupamiento, la sexta capa es una capa "completamente conectada" (idéntica a las capas utilizadas típicamente en redes neuronales más simples) y la séptima capa normaliza el resultado en diez nodos, de forma tal que cada nodo represente un número del 0 al 9, y el valor que tome ese nodo sea la probabilidad de que la imagen utilizada represente a cada número. Actualmente, las arquitecturas utilizan algunas operaciones diferentes, como los bloques *fused-MBConv* de las arquitecturas de la familia EfficientNetV2 (Tan and Le, 2019) que permiten entrenar más rápido y extraer patrones más complejos con un número factible de parámetros.

[7] La cantidad de épocas suele tratarse como hiperparámetro, es decir, que no existe un criterio objetivo de cuál es la cantidad óptima *a priori* para alcanzar el mejor desempeño del modelo, por lo que se selecciona empíricamente, probando diferentes alternativas y seleccionado la que produzca un mejor desempeño del modelo.

En las siguientes subsecciones se describen las definiciones necesarias para el entrenamiento de una red neuronal convolucional, como los conjuntos de Entrenamiento, Validación y Prueba (Sección IV-A), las métricas de evaluación (Sección IV-B) y la arquitectura y el conjunto de hiperparámetros (Sección IV-C). En el Anexo se desarrollan en detalle el proceso de selección de los hiperparámetros (A), la selección del tamaño de las imágenes (B), y la evolución de las métricas de entrenamiento (C).

IV-A. Separación en conjunto de entrenamiento, validación y prueba

Los modelos de redes neuronales generalmente presentan una tendencia al sobreajuste (*overfitting*), es decir, a aprender no solo las características generales del conjunto de entrenamiento, sino a extraer también sus características particulares, limitando la capacidad de aplicación a cualquier otro conjunto de datos. Dado que en este trabajo se busca construir un modelo que permita estimar el ingreso sobre cualquier conjunto de imágenes del Área Metropolitana de Buenos Aires, en varios años diferentes e incluso de áreas urbanas con características similares, el objetivo es que el modelo seleccionado tenga la capacidad de generalizar los patrones encontrados más allá del conjunto de entrenamiento.

Por ello, se siguió la práctica estándar de utilizar tres subconjuntos disjuntos de entrenamiento, validación y prueba (Prince, 2024). La división en conjuntos de entrenamiento, validación y prueba se realizó a nivel de radio censal y no a nivel de imagen. Por lo tanto, todas las imágenes de un determinado radio censal pertenecerán unívocamente al mismo conjunto. El conjunto de entrenamiento está comprendido por aproximadamente el 75 % de los radios censales del AMBA (6901 radios censales), el conjunto de validación está comprendido por el 5 % de los radios censales (493) y el conjunto de prueba es el restante 20 % de los radios censales (1798).

El criterio de asignación de cada radio censal a cada conjunto fue el siguiente. Para definir el conjunto de prueba se seleccionaron dos franjas de 0.05 grados de ancho (aproximadamente 4.5km) que cortan el AMBA de norte a sur y que contienen el 20 % de los radios censales. Específicamente, se seleccionó el área limitada entre las longitudes -58.71 y -58.66, y entre -58.41 y -58.36. Se descartaron aquellas imágenes que estuvieran atravesadas por las latitudes mencionadas, con el objetivo de evitar una posible contaminación cruzada entre los conjuntos de entrenamiento y prueba.

La Figura 6a muestra la distribución geográfica de los conjuntos de entrenamiento y de prueba. Por su parte, la Figura 6b compara la distribución del ingreso per cápita entre ambos conjuntos. Si bien las distribuciones no son idénticas, el conjunto de prueba presenta una mayor variabilidad que el conjunto de entrenamiento, por lo que resulta de utilidad para hacer una evaluación completa de la capacidad predictiva del modelo.

Para definir el conjunto de validación, los radios censales que no se utilizaron para el conjunto de prueba se asignaron aleatoriamente al conjunto de entrenamiento y validación, para cumplir con la distribución de 75 % entrenamiento, 5 % validación y 20 % prueba.

IV-B. Métricas de evaluación

Una vez definidos los radios censales que pertenecen al conjunto de prueba y el proceso de selección de las imágenes de este conjunto, es importante definir el mecanismo por el cuál se seleccionará el "mejor" modelo entre las diferentes arquitecturas utilizadas y dentro del mismo entrenamiento para una arquitectura en particular.

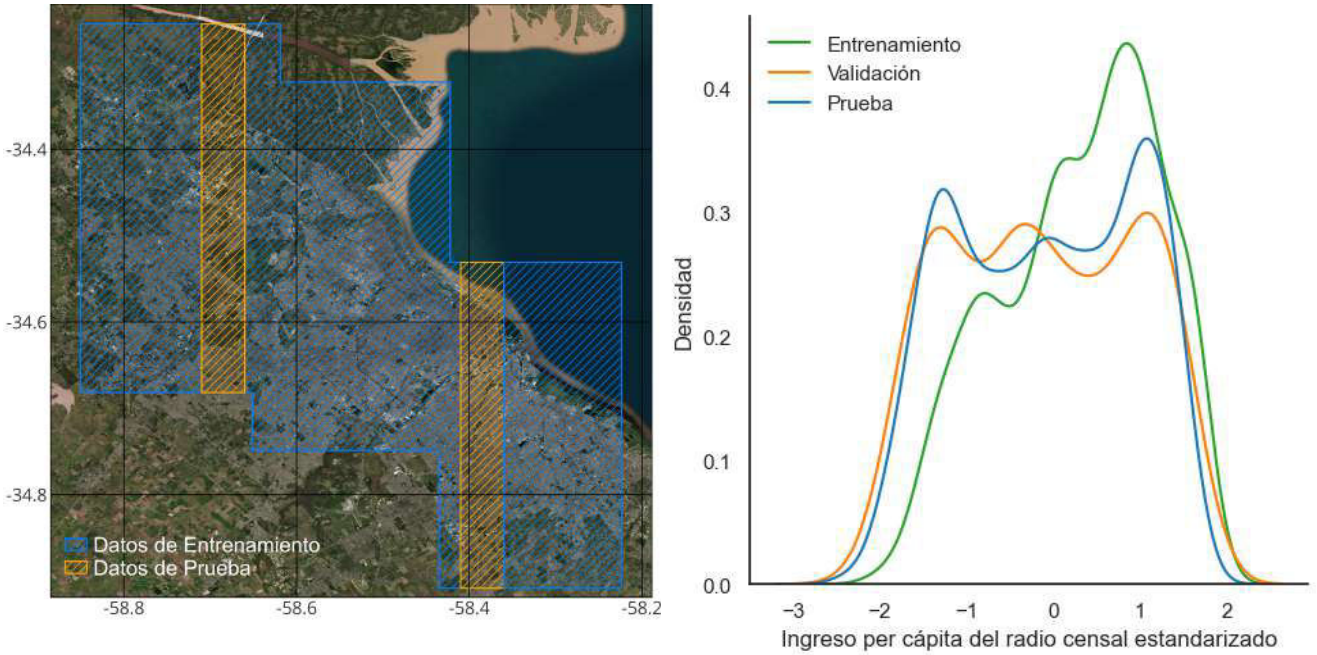
Se utilizó el Error Cuardático Medio (MSE) como métrica para entrenar y evaluar el modelo. En el entrenamiento, se utilizó esta métrica como función de pérdida, comparando la predicción del ingreso para cada una de las imágenes generadas en cada época, con el ingreso medio del radio censal. Sin embargo, para la evaluación se utilizó el MSE calculado a nivel de radio censal, en vez de a nivel de imágen. Esto se debe a que calcular el MSE de forma directa sobre cada una de las imágenes implicaría una sobreestimación del verdadero error de la estimación, ya que parte de la diferencia entre la predicción y el ingreso de ese radio censal puede deberse a la variación dentro de un mismo radio censal. En otras palabras, un modelo que prediga correctamente el ingreso a partir de imágenes tendría un MSE alto en muchos de los radios censales, debido a que la predicción se corresponde a un subconjunto del radio censal (el área delimitada por la imágen) y no al ingreso medio de la totalidad de ese radio. La alternativa propuesta es computar el MSE sobre la media de las predicciones para cada radio censal, ya que, de esta forma, se estaría comparando la estimación el ingreso medio estimado mediante todas las imágenes de ese radio, contra el ingreso medio de cada radio censal. Esta métrica, que calcula el MSE promediando las predicciones de cada radio censal, se utiliza para calcular el error en los conjuntos de validación y prueba. Para simplificar la comparación, en la sección de resultados se calcula el R^2, que normaliza el MSE por la varianza de las observaciones, generando un indicador que va de 0 a 1, siendo 0 un ajuste nulo (MSE igual a la varianza) y 1 un ajuste perfecto (MSE igual a cero).

Formalmente, la sobreestimación del error cuadrático medio computado sobre la totalidad de las imágenes puede observarse si se considera la siguiente expresión:

$$MSE = \frac{1}{n_j n_h} \sum_{j \in J_T} \sum_{h \in H_j} \left(\hat{p}_{hj} - \hat{Y}_j \right)^2 \tag{4}$$

donde $\hat{p}_{hj}$ representa la predicción del modelo de redes neuronales de la imagen h que pertenece al radio censal j, $\hat{Y}_j$ representa el ingreso per cápita del radio censal j, J_T representa el conjunto de radio censales que pertenecen al grupo de prueba y H_j el conjunto de imágenes que pertenecen al radio censal j.

Siguendo la lógica postulada, en un modelo estimado correctamente $\hat{p}_{hj}$ debería presentar una variación dentro del radio censal j, mientras que $\hat{Y}_j$ es constante dentro del mismo radio. Por lo tanto, la métrica de evaluación subestima la precisión del modelo. Para obtener un error cuadrático medio apropiado, en vez de comparar cada una de las imágenes, es



(a) Distribución geográfica de las imágenes.

(b) Comparación del ingreso a nivel de radio censal.

Figura 6: Distribución del conjunto de entrenamiento y prueba.

posible comparar la media de las predicciones de todas las imágenes dentro del radio censal con el ingreso per cápita a partir de los datos de los hogares. Siguiendo esto, el **error cuadrático medio** utilizado para evaluar el desempeño del modelo será

$$MSE^{test} = \frac{1}{n_j} \sum_{j \in J_T} \left(\hat{P}_j - \hat{Y}_j \right)^2 \tag{5}$$

que se calcula sobre todos los radios censales j que pertenezcan al subconjunto de prueba J_T y donde $\hat{P}_j$ representa la media de las predicciones de todas las imágenes que componen el radio censal j.

$$\hat{P}_j = \frac{1}{n_h} \sum_{h \in H_j} \hat{p}_{hj} \tag{6}$$

Dado que el valor del ingreso $\hat{Y}_j$ se compone como la media de las predicciones del ingreso per cápita de los hogares dentro de un radio censal j, se aplicará una lógica similar para calcular la predicción media del modelo. Para cada radio censal, se constuye una grilla de imágenes del mismo tamaño que las utilizadas en el conjunto de entrenamiento; es decir, se generan predicciones para toda el área que se corresponde al radio censal. Específicamente, para construir la grilla se consideran únicamente las imágenes cuyo centroide se encuentre dentro del radio censal, por lo que algunas imágenes se extienden levemente sobre otros radios.

La Figura 7 presenta un ejemplo de radio censal sobre el cuál se construye la grilla para generar las predicciones. La predicción media $\hat{P}_j$ del radio censal j se constuye, entonces, computando la media de las imágenes de la grilla construida.

Figura 7: Construcción de grilla de imágenes de 128x128px (50x50mts) para calcular la predicción media del ingreso en un radio censal.

Fuente: Imagen satelital provista por la CONAE.

IV-C. Arquitectura y configuración del Modelo

Como arquitectura base de la red se utilizó una versión de la red neuronal convolucional EfficientNetV2 (Tan and Le, 2019). Estos modelos son capaces de lograr un alto rendimiento con una cantidad relativamente baja de parámetros. Específicamente, se utilizó una red EfficientNetV2-S, con una tasa de aprendizaje de 0.0001. Como optimizador se utilizó el método de estimación de momento adaptativo acelerado de Nesterov (Dozat, 2016). El entrenamiento se realizó a lo largo de 150 épocas y se seleccionaron los parámetros

de la época con menor Error Cuadrático Medio sobre el conjunto de validación a nivel de radio censal (un 5 % de los radios censales del conjunto de entrenamiento). El modelo se entrenó utilizando una GPU GeForce RTX 2060 Super, y el entrenamiento del modelo requirió aproximadamente 52 horas. En el Anexo se desarrollan en detalle el proceso de selección de los hiperparámetros (A), la selección del tamaño de las imágenes (B), y la evolución de las métricas de entrenamiento (C).

V. Construcción de un mapa del ingreso a nivel de grilla

Una vez entrenado el modelo, es posible obtener una estimación del indicador de interés a partir de una imagen del tamaño utilizado, sin proveer más información que la contenida en ella. Como el indicador utilizado es el ingreso per cápita de los hogares, es posible generar un mapa altamente desagregado de la distribución espacial del ingreso.

Para generar estas predicciones es necesario construir una grilla que contenga la totalidad del área de interés, donde cada celda será una imagen sobre la cuál se realizará la predicción del ingreso. Como las imágenes de entrenamiento cubren un área de 50×50 metros, la grilla de predicciones estará compuesta por celdas de ese tamaño. Una vez definida esta grilla, se generan las predicciones de la totalidad de las imágenes, tanto las pertenecientes al grupo de entrenamiento como el de prueba, para las imágenes disponibles en 2013, 2018 y 2022.

La construcción de estos mapas permitirá generar estimaciones del ingreso para celdas de 50×50 metros, cuadruplicando la resolución del indicador en los radios censales pertenecientes a las zonas de mayor densidad poblacional, y mejorándola muchas veces más en zonas más alejadas del centro de la ciudad. Esto es posible ya que la selección de las imágenes para cada radio censal fue aleatoria. Como las redes neuronales convolucionales son robustas a errores no sistemáticos en los datos de entrada (Burke et al., 2021), se espera que el modelo asigne un peso menor en el entrenamiento a los datos "incongruentes" con el patrón general. Si esto ocurre, entonces el modelo puede ser aplicado a todos los radios censales, generando estimaciones consistentes. En segundo lugar, se podrán construir series de tiempo para cada una de las celdas de la grilla generada en el punto anterior. Esto es posible ya que para el algoritmo es idéntico analizar imágenes de diferentes momentos del tiempo. Entonces, las subsecuentes predicciones en el tiempo para una imagen del indicador de interés permitirá construir mapas del indicador a lo largo del tiempo.

Es necesario destacar que este enfoque solo podrá detectar cambios estructurales en las regiones, determinados por la calidad de las viviendas, el nivel de urbanización o de vegetación, entre otras características. Lógicamente, no podrá detectar cambios en el ingreso generados por la coyuntura económica, ya que el modelo nunca accederá a esa clase de información. Sin embargo, una herramienta como esta es sumamente útil para identificar espacialmente a las zonas donde viven las personas menos favorecidas económicamente, así como también las personas con mayores recursos económicos. A su vez, permite identificar cambios habitacionales relevantes, como nuevos desarrollos urbanos y/o cambios importantes en la distribución espacial de la riqueza.

VI. Resultados

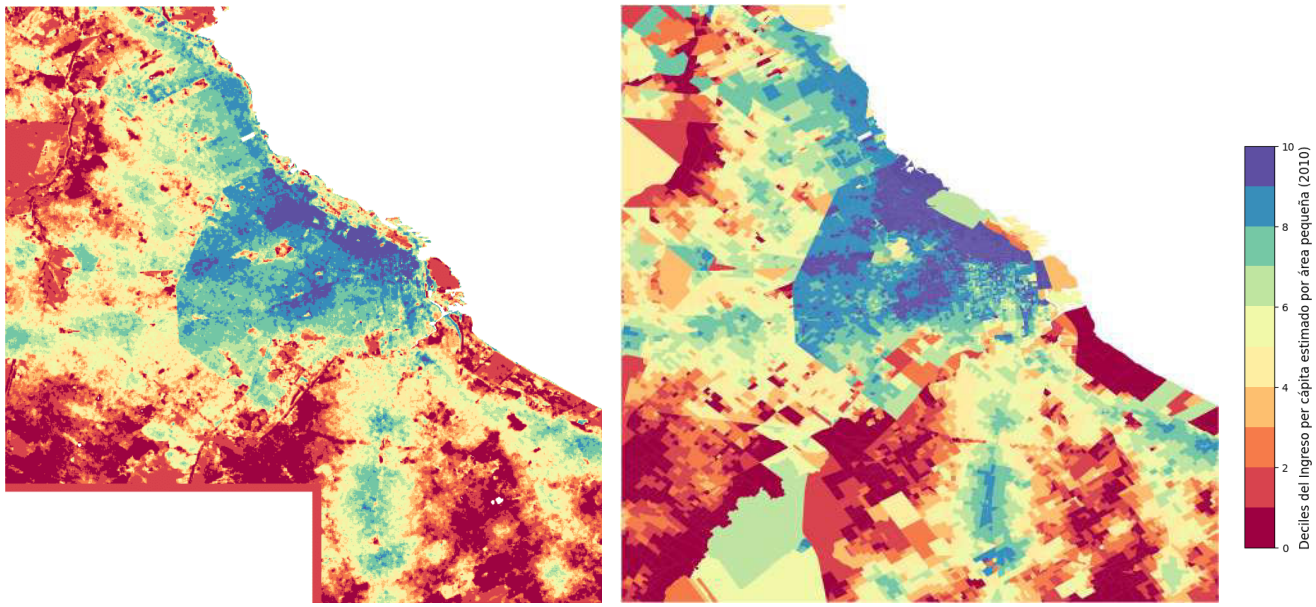
La red neuronal entrenada permite generar predicciones de muy alta resolución del ingreso per cápita. El modelo genera estimaciones del ingreso sobre una grilla de celdas de 50×50 metros. Para la Argentina, no existen datos del ingreso per cápita con este nivel de resolución. Por ello, para evaluar la precisión del modelo, es necesario considerar diferentes niveles de agregación, para verificar si el modelo genera predicciones consistentes con los datos disponibles del ingreso para el Área Metropolitana de Buenos Aires, así como también verificar visualmente que las predicciones del modelo tengan una correspondencia con lo observado en las imágenes. En la sección VI-A, se comparan las predicciones del modelo con el ingreso per cápita de cada radio censal. La sección VI-B analiza cualitativamente —por la falta de datos en ese nivel de resolución— las predicciones del modelo al interior de los radios censales. La sección VI-C compara las predicciones agregadas (i) para la totalidad del ABMA y (ii) a nivel municipal/comunal, con los datos de la Encuesta Permanente de Hogares y las estimaciones del ingreso por área pequeña para el censo 2010. Finalmente, la sección VI-D evalúa la validez de las predicciones del modelo a través del tiempo, comparando los resultados para 2013, 2018 y 2022.

VI-A. Una estimación espacial del ingreso

En esta sección se presenta el principal resultado de este trabajo: un mapa que estima, con una resolución de 50 metros por 50 metros, el logaritmo del ingreso per cápita promedio de los hogares que viven dentro de cada celda. La Figura 8a presenta las predicciones para la grilla de imágenes satelitales de 2013. A su derecha, la Figura 8b presenta los datos en su forma original, es decir, la estimación de área pequeña del ingreso generada utilizando los datos del censo. El ingreso se expresa en deciles del año 2010, para facilitar su interpretación. Como resulta evidente, la principal ventaja de esta metodología es que permite aumentar la granularidad de la estimación del ingreso, al utilizar la información disponible de forma no estructurada en las imágenes satelitales.

El modelo seleccionado tiene un muy buen desempeño predictivo, alcanzando un R^2 de 0.878 sobre el conjunto de prueba. En otras palabras, el modelo logra capturar más del 87 % de las variaciones espaciales del ingreso sobre un conjunto de imágenes que el modelo nunca vió previamente.

La importante capacidad predictiva del modelo se hace evidente en la Figura 9, que compara para cada radio censal del conjunto de prueba el logaritmo del ingreso per cápita estimado por área pequeña con el valor predicho por la red neuronal. Ambas variables están estandarizadas para tener media cero y desvío estándar igual 1. El modelo predice con un muy bajo nivel de error los radios censales con ingresos más bajos, en particular aquellos por debajo del ingreso medio (cuadrante inferior izquierdo). En ingresos medios y altos, el



(a) Predicciones utilizando la estrategia propuesta. (b) Datos originales del ingreso en el AMBA.

Figura 8: Estimación espacial del ingreso per cápita.

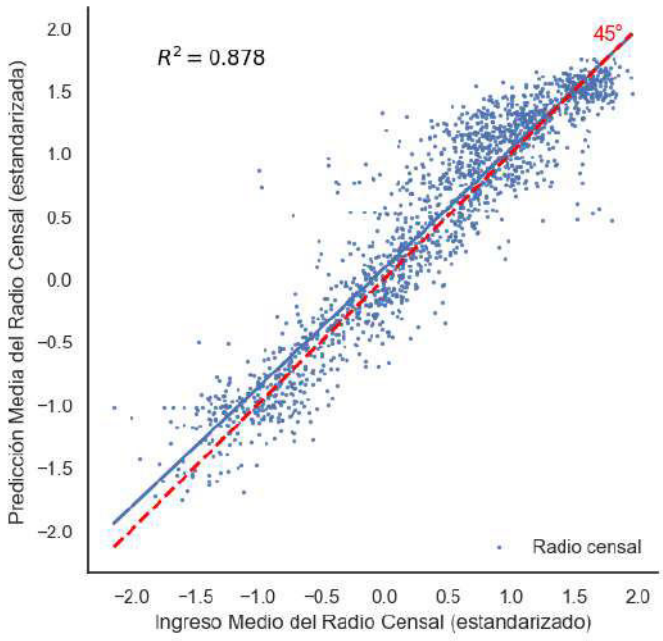


Figura 9: Comparación del ingreso predicho y observado por Radio Censal, sobre el conjunto de prueba.

modelo presenta una mayor variabilidad en las predicciones. Este resultado es razonable, ya que los hogares de bajos ingresos tienen menos posibilidades de elección del tipo de vivienda, por lo que la variabilidad observada mediante las imágenes satelitales es menor.[8]

La Tabla I compara el resultado obtenido por el modelo utilizado con metodologías similares implementadas en la literatura. Los resultados presentados implican una mejora sustantiva respecto al desempeño de modelos similares, ya que utilizan una menor resolución espacial y obtienen un desempeño mejor que otras estrategias. Es importante destacar que la precisión de los diferentes modelos no puede compararse de forma directa, al tratarse de bases de datos diferentes y metodologías alternativas aplicadas a zonas diferentes y tamaños de imágenes distintos, tanto en píxeles como en metros. A pesar de esto, la Tabla I ofrece un panorama muy claro respecto a la utilidad del modelo en comparación a la literatura, ya que genera resultados precisos sobre un tamaño de imágenes 20 veces menor al utilizado en otros trabajos relacionados.

VI-B. Evaluación ingreso per cápita estimado a nivel de grilla

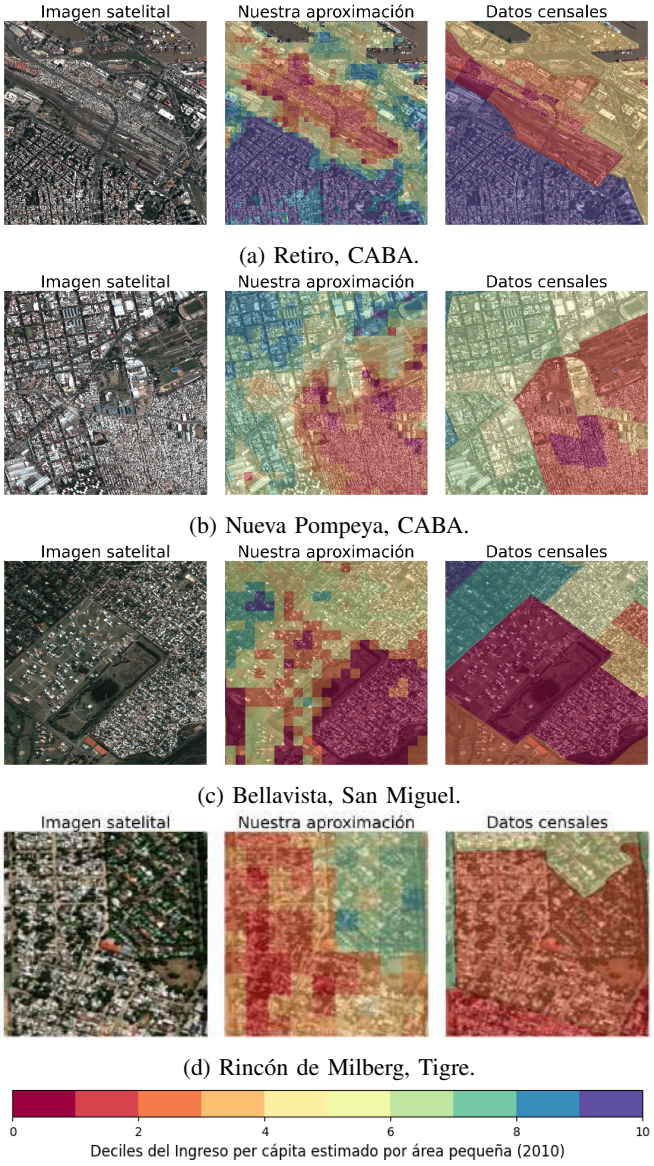
Dada la falta de datos a nivel sub-radio censal disponibles para evaluar la precisión de las estimaciones a nivel de grilla, en esta sección se evalúan cualitativamente las predicciones del modelo. En la Figura 10 se presentan una serie de casos testigo, buscando destacar situaciones complejas que desafíen la capacidad predictiva del model, como barrios cerrados de altos ingresos aledaños a asentamientos informales. Todos los casos se encuentran dentro de las franjas que delimitan al conjunto de prueba, por lo que no se trata de imágenes con las que el modelo haya sido entrenado.

Resulta interesante que, en los casos presentados, los asentamientos informales como el Barrio 31 (en Retiro) y el 21-24 (en Nueva Pompeya) son capturados correctamente por el modelo, delimitando el área donde a simple vista hay viviendas informales. Asimismo, las zonas con edificios o casas de mejores materiales son clasificados con mayores ingresos, en línea con los datos del censo. En las zonas no pobladas,

[8]Intuitivamente, un hogar de ingresos medios o altos puede elegir vivir en un barrio menos costoso y dedicar esos ingresos disponibles a otros consumos, mientras que un hogar con ingresos bajos no tiene esa posibilidad.

Tabla I: Comparación del modelo propio con estrategias de la literatura.

	Modelo propio	Piaggesi et al (2019)	Rolf et al (2021)	Khachiyan et al (2022)	Henderson et al (2012)
Modelo	EfficientNetV2	ResNet50	Custom CNN	Custom CNN	Lineal
Variable de interés	Ingreso per cápita	Ingreso per cápita	Ingreso per cápita	Ingreso total	Crecimiento económico
Imágenes utilizadas					
Tipo	Reflactancia Diurna	Reflactancia Diurna	Reflactancia Diurna	Reflactancia Diurna	Luces Nocturnas
fuente	Pléiades Neo	DigitalGlobe/GMaps	GMaps	LandSat 8	Suomi NPP/NOAA-20
tamaño	50x50m	1x1km	1x1km	1.2x1.2km	Media por país
en pixeles	128x128	224x244	256x256	128x128	-
R^2	0.878	0.691	0.45	0.749	0.769



(a) Retiro, CABA.

(b) Nueva Pompeya, CABA.

(c) Bellavista, San Miguel.

(d) Rincón de Milberg, Tigre.

Figura 10: Comparación de datos censales con predicciones de la red neuronal para casos testigo del conjunto de prueba.

Fuente: Imagenes satelitales provistas por la CONAE.

como las vías de tren, el puerto o los espacios verdes, las predicciones son aleatorias y no siguen ningún patrón de interés para este trabajo.

A su vez, como se observa en el caso de Bellavista, algunas zonas con incipiente urbanización —como el barrio con piletas en construcción que se observa al centro de la imagen— son clasificadas como de ingresos medios bajos, en contraste con los datos censales que asignan ingresos bajos a esa área. Esto puede deberse a que ese radio censal contiene muchas observaciones de una sección relativamente pobre, y que además contiene unos pocos hogares de ingresos más altos. En la sección VI-D se evalúa este caso en comparación con 2018, donde se destaca que el barrio que en 2013 está en construcción ya se encuentra establecido, y en consecuencia el modelo predice ingresos altos para esa sección.

Finalmente, la Figura 10d muestra un caso en el que un mismo radio censal de la ciudad de Tigre contiene parte de un barrio cerrado, típico de hogares de altos ingresos, y un asentamiento informal, característico de hogares de ingresos bajos. El radio censal, en promedio, tiene un ingreso medio-bajo, ya que promedia el ingreso de todos los hogares. Como se observa en la Figura 10d, el modelo logra capturar mejor la distribución espacial del ingreso, ya que logra asignar ingresos medios-altos a las celdas del barrio cerrado e ingresos bajos al asentamiento informal. En un caso como este, si se compara cada imagen contra el ingreso promedio del radio censal, el error será mucho mayor que el calculado comparando el ingreso promedio de las predicciones contra el ingreso medio del radio censal.

VI-C. Predicción de Indicadores Económicos en 2010

Como se mencionó, el principal objetivo del modelo planteado es lograr capturar la distribución espacial del ingreso con una resolución espacial muy alta. Sin embargo, resulta interesante evaluar si, al agregar las predicciones de la grilla a niveles geográficos más altos, el modelo genera estimaciones precisas de algunos indicadores económicos relevantes. Por lo tanto, se comparan las predicciones del modelo con indicadores observados en la Encuesta Permanente de Hogares y en la estimación de área pequeña del Censo, para el año 2010. Idealmente, el modelo generará predicciones consistentes tanto con los datos de entrenamiento como con los datos observados de la Encuesta Permanente de Hogares. En términos generales, el modelo captura algunas dinámicas de estos indicadores, con ciertas limitaciones. En esta sección y, en particular, en la sección VII, se discute por qué el modelo planteado podría no estar capturando estas dinámicas.

La Tabla II compara estadísticas descriptivas de la distribución del ingreso (1) de los 5,348 hogares observados en la Encuesta Permanente de Hogares del 2do semestre de 2010, (2) de la predicción del ingreso para los 13,491 radios censales del Censo en el AMBA, es decir, la estimación del ingreso por *área pequeña* para cada radio censal (ver ecuación 3) y (3) la predicción del ingreso para las 539,809 celdas del AMBA.

En los casos (2) y (3), donde las observaciones no son hogares sino áreas, los indicadores se calcularon ponderando cada área por la cantidad de hogares que viven en ella. Para los datos del censo la información está directamente disponible para cada radio censal. Para las predicciones a nivel de grilla, se utilizaron las estimaciones de población de *Gridded Population of the World* (Center for International Earth Science Information Network (CIESIN), 2018), que estima la población en una grilla con una resolución de 30-arcosegundos (aproximadamente 1km). Esto se realiza para reducir la ponderación de zonas que tienen baja o nula densidad poblacional, como pueden ser algunas áreas de las afueras del AMBA. De no realizar este filtro, la estimación del ingreso medio sería excesivamente baja, subestimando el ingreso medio "real".

Se comparan los siguientes indicadores de la distribución del ingreso: el ingreso medio y mediano, que dan una idea de la tendencia central de la distribución; el Ratio 90-10 y 50-10, dos indicadores de desigualdad económica que comparan el ingreso del hogar que se encuentra en el percentil 90 (50) con el ingreso del hogar que se encuentra en el percentil 10; el coeficiente de Gini, un indicador de desigualdad que compara el ingreso de todos los hogares de la distribución; y la tasa de incidencia de la pobreza utilizando la línea nacional de pobreza (FGT(0)).

Se destacan varias conclusiones de la Tabla II. En primer lugar, las medidas de tendencia central muestran una subestimación respecto de los datos observados en la EPH a medida que se avanzan en los pasos para la estimación. Esto es particularmente cierto para la media, que pasa de un ingreso per cápita promedio de los hogares observado de $820.94 a un ingreso medio predicho de $649.37. Es posible explicar esto si se tiene en cuenta que la heterogeneidad en las viviendas, observada por el modelo, puede ser menor a la heterogeneidad en los ingresos. Por ejemplo, esto ocurriría si la elasticidad-ingreso de la demanda de viviendas es menor a 1, fenómeno que se observa, por ejemplo, en Estados Unidos (Hansen et al., 1998) y en México (Adamuz Peña and González Tejeda, 2016). Además, cuando se realizan predicciones de un modelo que busca minimizar el error cuadrático medio —como una regresión lineal, en la ecuación 1— la dispersión de las predicciones se ve significativamente comprimida, reduciendo los indicadores de desigualdad y variabilidad. En una distribución asimétrica, como el ingreso, es esperable también que ante esta caída de la variabilidad la media se reduzca, al reducir el peso que ingresos extremos tienen sobre este indicador. En la Figura 9 la suma de ambos efectos se ven claramente, si se considera que, mientras que el ingreso de los radios censales con datos del censo la gran mayoría de las observaciones se encuentran entre -2 y 2 desvíos estándar de la media, la gran mayoría de las predicciones está entre -1.5 y 1.5.

Tabla II: Comparación de indicadores económicos.

	(1) Observado EPH	(2) Censo (estimación)	(3) Img. Satelital
Unidades	Hogares	Radios Censales	Celdas
Observaciones	5,348	13,491	539,809
Año	2010	2010	2013
Indicadores:			
Media	820.94	691.84	649.37
Mediana	611.45	567.16	541.75
Ratio 90-10	8.05	3.59	3.78
Ratio 50-10	3.01	1.62	1.74
Gini	0.426	0.275	0.282
FGT(0)	32.17	30.30	37.01

Por ambas razones, es también esperable que las medidas de desigualdad se subestimen, como se destaca comparando los indicadores para los modelos (2) y (3) respecto a los datos observados en la EPH, columna (1). Esta subestimación tan importante en los indicadores de pobreza y desigualdad se explican, principalmente, al pasar de los datos de la columna (1) a la columna (2), es decir, al estimar por área pequeña el ingreso de los radios censales. Por el contrario, la red neuronal (columna (3)) genera predicciones muy similares a los datos con los cuales fue entrenado (columna (2)). De hecho, es posible que la red neuronal capture variaciones no incorporadas en el modelo de área pequeña, lo que explicaría por qué las medidas de desigualdad son más similares si se comparan las columnas (3) y (1), que si se comparan las columnas (2) y (1). Esto es particularmente destacable si se tiene en cuenta que, para generar las predicciones de la red neuronal, los datos utilizados para el entrenamiento fueron los de la columna (2). En otras palabras, el modelo generado logra capturar variaciones en el ingreso espacial que no estaban incorporados en los datos originales. Por otra parte, la pobreza no se logra estimar correctamente por ser altamente sensible a la forma de la cola izquierda de la distribución del ingreso, quedando muchos más radios censales por debajo de la línea de pobreza que los observados en la EPH y los estimados sobre el censo.

Cuando se analiza el ingreso agregado a nivel municipal (o comunal), las conclusiones son cualitativamente muy similares, pero mucho más precisas. La Figura 11 compara, para cada municipio del Área Metropolitana de Buenos Aires[9], diferentes indicadores económicos calculados con el indicador del ingreso estimado por área pequeña[10], con los mismos indicadores computados utilizando ingreso predicho por la red neuronal para cada una de las celdas de la grilla. De la Figura se extraen tres conclusiones principales. En primer lugar, a nivel municipal, las estimaciones por imágenes satelitales subestiman sistemáticamente las medidas de tendencia central del censo, lo que explica por qué el modelo subestima la desigualdad y sobrestima la pobreza. En segundo lugar, el modelo estima correctamente la tendencia de todos los indicadores a nivel municipal, ya que, por ejemplo municipios con mayor desigualdad con los datos del censo presentan una

[9]Dentro de la Ciudad de Buenos Aires, la desagregación se realiza a nivel de Comuna.

[10]Es decir, el indicador generado en la ecuación 3 y que se corresponde a la columna (2) de la Tabla I.

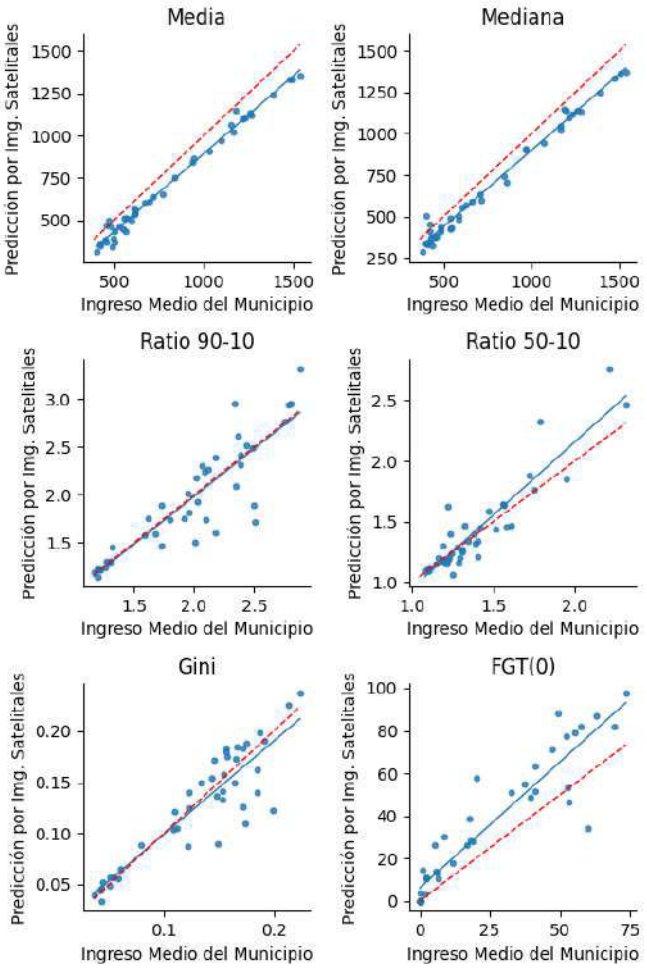


Figura 11: Comparación de indicadores a nivel Municipal.

Nota: cada municipio del AMBA es representado como un punto.

mayor desigualdad estimada utilizando imágenes satelitales. Finalmente, es destacable que las estimaciones de las medidas de tendencia central son por mucha diferencia las más precisas, generando un indicador que subestima el ingreso pero estima muy correctamente el ordenamiento de los municipios. En otras palabras, el modelo captura correctamente (a nivel municipal) la distribución espacial del ingreso, aunque no logra capturar correctamente el nivel de ingreso per cápita de los hogares.

VI-D. Predicciones del ingreso en el tiempo

Utilizando el mismo modelo entrenado que en las secciones previas, es posible generar una grilla de predicciones del ingreso para imágenes de 2018 y 2022. Para ello, simplemente hay que generar las predicciones para cada una de las imágenes de la grilla.

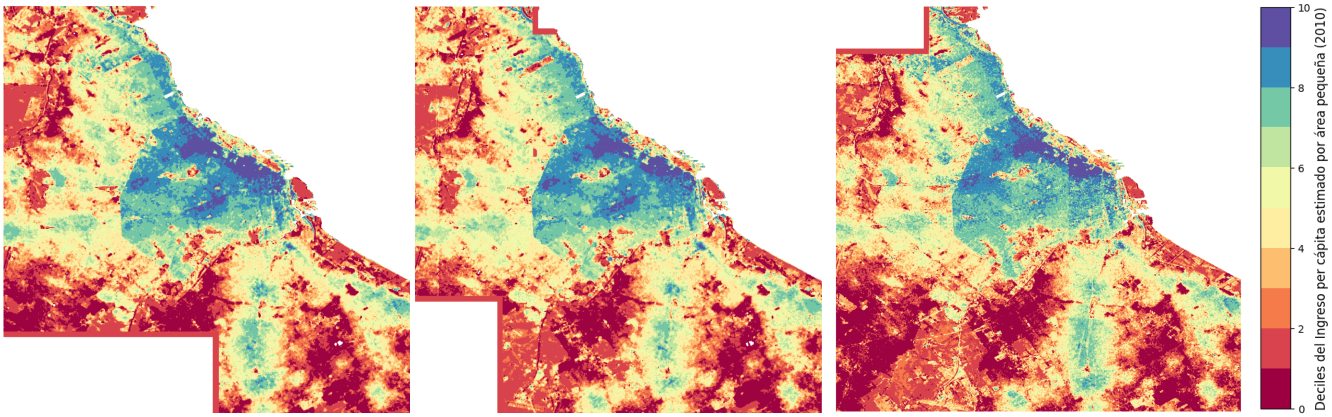
La Figura 12 muestra la evolución del ingreso per cápita, expresado en deciles del año 2010, para los tres años utilizados. A la izquierda, se replica la Figura 8a, para comparar más fácilmente con las imágenes de los años posteriores. En líneas generales, el modelo predice una distribución del ingreso espacial muy similar para los tres años, lo que destaca la consistencia del modelo al predecir en imágenes de diferentes años. Por otra parte, para los años 2018, y en particular para 2022, se observa una menor correlación espacial entre celdas aledañas. Esto puede verse en detalle en las zonas donde limitan celdas que pertenecen a diferentes deciles. Mientras que en las imágenes de 2013 se observan *grupos* de varias celdas donde todas pertenecen al mismo decil, en las imágenes de 2022, para una misma zona, las predicciones son más variadas. En la sección VII se discute brevemente por qué esto puede estar ocurriendo.

A nivel de grilla, las predicciones del modelo parecen predecir apropiadamente la evolución en la infraestructura urbana a través del tiempo. A modo de ejemplo, se ofrece uno de los casos presentados en la sección VI-B, correspondiente a la localidad de Bella Vista, San Miguel. La Figura 13 muestra la evolución del indicador entre 2013 y 2018 para una imagen correspondiente al conjunto de prueba, por lo que no fue previamente vista por el modelo.[11] Como puede observarse en el lado izquierdo de las imágenes, existe un área de incipiente urbanización en 2013 que para 2018 ya se encuentra desarrollada. En 2013, muchas de las casas están en construcción, muchas sin techos, por lo que el modelo predice ingresos medios y bajos; las zonas donde existen construcciones sin terminar y/o abandonadas, típicamente son habitados por hogares de menores recursos. Para 2018, la urbanización está más definida, con casas grandes, acompañadas con piletas y grandes jardines, lo que típicamente identifica una zona habitada por hogares de ingresos más altos. El modelo logra capturar esa dinámica, al tiempo que las zonas aledañas que no presentaron cambios se mantienen relativamente constantes.

Así como en las secciones previas se estimaron indicadores económicos para el año 2010, al replicar la estimación de las grillas del ingreso para 2018 y 2022 es posible evaluar la evolución de estos indicadores a través del tiempo. Como es esperable, estos resultados tienen sus limitaciones cuando se comparan con el ingreso observado en encuestas como la EPH, ya que el ingreso corriente puede verse sujeto a variaciones de corto plazo, mientras que buena parte de las características observables en las imágenes satelitales se corresponden con variaciones del ingreso de medio o largo plazo.

La Figura 14 muestra la variación porcentual de diferentes indicadores del ingreso, tanto con el modelo aplicado a imágenes satelitales del año 2018 y 2022, como calculados con las EPH del semestre correspondiente a esas imágenes. Si bien la tasa de crecimiento de los indicadores no es capturada perfectamente por el modelo —en todos los casos las líneas se encuentran a algunos puntos de distancia— es destacable que para todos los indicadores la tendencia del indicador es capturada mediante las imágenes satelitales. Solo en unos pocos casos, como en el Ratio 50-10 y el coeficiente de Gini para 2018, el modelo predice incorrectamente la evolución temporal del indicador. Además, es importante destacar que los indicadores predichos por el modelo subestiman sistemáticamente su evolución en el tiempo. Esto se encuentra en línea

[11] No se cuentan con imágenes de esa sección para el año 2022 como para mostrar su evolución en ese año.



(a) 2013 (b) 2018 (c) 2022

Figura 12: Estimación espacial del ingreso per cápita en diferentes años.

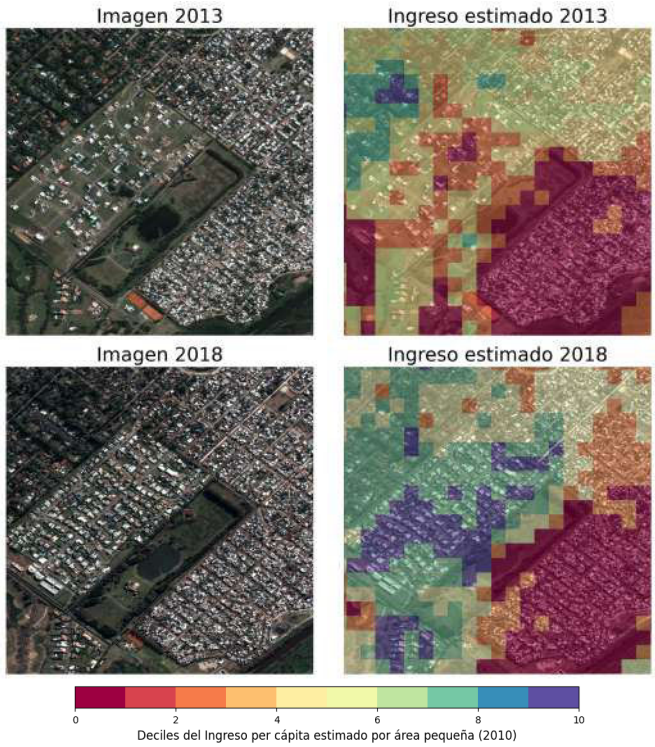


Figura 13: Detección de nuevas urbanizaciones en Bella Vista, San Miguel.

Fuente: Imagenes satelitales provistas por la CONAE.

con la interpretación de que el modelo no captura el ingreso corriente de las personas, sino una dimensión del ingreso de largo plazo. Esto se desarrolla en detalle en la sección VII.

VII. Discusión

A lo largo de la sección de Resultados se menciona que el indicador que predice el modelo es el (logaritmo del) ingreso per cápita familiar promedio de los hogares que habitan en el cuadrante de la imagen. Si bien los datos de entrenamiento efectivamente miden esa variable (ver Figura 3), es posible que las predicciones del modelo representen algo levemente diferente. Como el modelo utiliza únicamente información *visual* de las imágenes para la predicción, el modelo podrá capturar únicamente variaciones del ingreso que puedan observarse en mejoras materiales observables.

La interpretación propuesta en este trabajo es que el modelo captura una variable latente, la esperanza temporal del ingreso de los hogares. En otras palabras, el modelo captura el ingreso "de largo plazo" o el ingreso "promedio" a través del tiempo de los hogares. Si un hogar presenta cambios transitorios en el ingreso, pero puede mantener su vivienda en las mismas condiciones, el modelo no podrá capturar tal variación. Por el contrario, si el hogar no puede mantener su vivienda, o esta empieza a deteriorarse, es posible que las imágenes capturen tal variación.

Por otra parte, en el entrenamiento fue evidente que las predicciones del modelo se ven afectadas por cuestiones coyunturales no necesariamente relevantes para predecir el ingreso. Por ejemplo, al utilizar únicamente imágenes de 2013 para el entrenamiento —para el modelo final se utilizó una combinación de las imágenes de 2013, 2018 y 2022—, al predecir sobre las imágenes de 2018 el modelo predecía sistemáticamente menores ingresos, debido a que las imágenes provenían de un mes más seco, con la vegetación más amarilla. Como esa característica, manteniendo todo lo demás constante, se correlaciona con hogares de ingresos más bajos, el modelo incorrectamente predecía una caída en el ingreso. De la misma forma, se observaba un salto en el ingreso sobre las imágenes de 2022 debido a que las imágenes fueron capturadas sobre el atardecer. Esto produjo que todas las edificaciones tuvieran más sombras, hecho que, manteniendo lo demás constante, es un indicador de edificios de mayor altura y por lo tanto mayores ingresos en promedio. Este tipo de factores que pudieron ser detectados y resueltos combinando imágenes de diferentes años, es posible que presente nuevas mejorías utilizando un mayor número de imágenes de diferentes momentos del tiempo.

Otro hecho que cabe mencionar es la dificultad que tiene el modelo para predecir los indicadores económicos tanto en el año base como a través del tiempo. Una solución lógica a tal

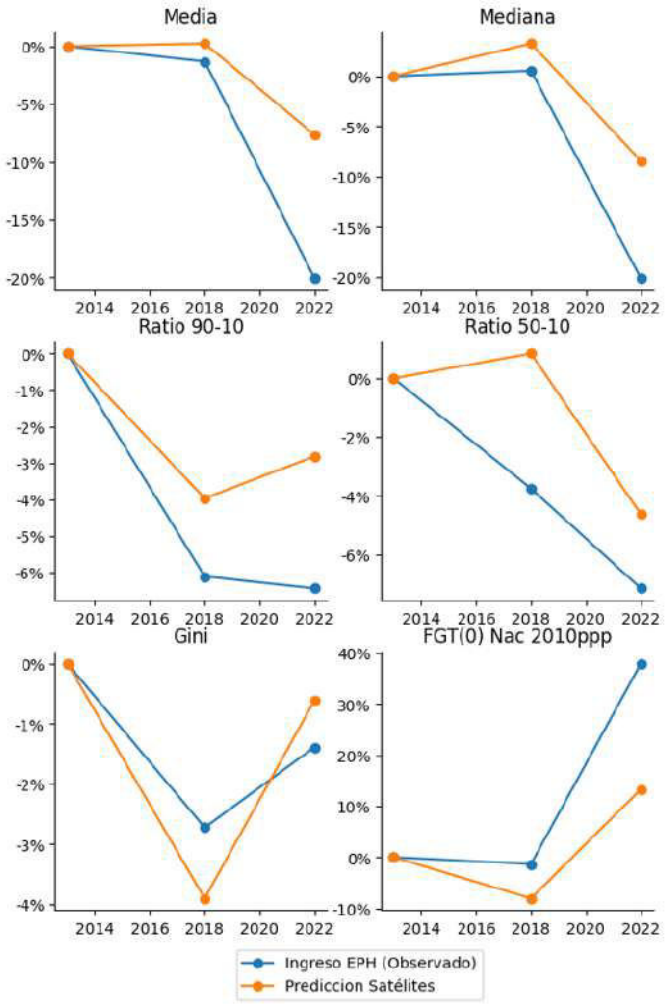


Figura 14: Indicadores económicos estimados para el AMBA a través del tiempo.

situación es implementar una estrategia similar a la utilizada por Rao and Molina (2015) sobre las estimaciones de área pequeña: agregar indicadores de encuestas a las estimaciones *ad-hoc* como la presentada aquí. Por ejemplo, Khachiyan et al. (2022) agregan indicadores como el ingreso medio de cada año a la última capa de la red neuronal, y logran una mejora sustancial sobre el rendimiento del modelo.

Un punto relevante a tener en cuenta es que el rendimiento actual del modelo, a nivel de radio censal, se compara únicamente con los datos generados utilizando técnicas de área pequeña. Uno de los problemas que presenta esta técnica es que, al depender de un modelo de regresión lineal, las variables observables utilizadas tienen un poder limitado para explicar variaciones sobre la distribución del ingreso. En este sentido, cualquier mejora en los datos de entrada implicará una mejoría directa en el modelo planteado.

Finalmente, un problema adicional que presenta la metodología en su estado actual es que el modelo no permite identificar si existen hogares viviendo en la celda sobre la cuál se predice el ingreso. Esto genera en una problemática marcada, ya que extender esta metodología a zonas de menor población o incluso áreas con amplias extensiones no habitadas implicaría ingresar al modelo información irrelevante. En subsecuentes iteraciones se buscará construir un modelo de dos etapas, que primero prediga si existe al menos una vivienda en la imagen, y posteriormente, si esta existe, generar la predicción del ingreso. De la misma forma, podría realizarse una estimación conjunta del ingreso con, por ejemplo, la densidad poblacional de la imagen, para resolver tal cuestión. La dificultad para implementar cualquiera de estas estrategias es la falta de datos de entrenamiento al nivel de resolución utilizado, al menos para el Área Metropolitana de Buenos Aires.

VIII. Conclusiones

En este trabajo se propone una metodología para desarrollar un modelo de predicción del ingreso per cápita a partir de imágenes satelitales. Dentro de esta metodología, se incluye una estrategia para la generación de una base de datos aplicable a una red neuronal, utilizando datos censales georreferenciados en polígonos irregulares, en conjunto con datos de encuestas que proveen la información del ingreso de los hogares. Se utilizó como arquitectura del modelo la red EfficientNetV2 (Tan and Le, 2019). Los resultados demuestran que el modelo alcanza una alta precisión en la predicción del ingreso per cápita, superando la resolución espacial de metodologías alternativas presentadas en la literatura.

La aplicación de este modelo permite construir un mapa del ingreso per cápita del Área Metropolitana de Buenos Aires, Argentina, para los años 2013, 2018 y 2022, a una escala espacial mucho más detallada que la presentada en los datos originales, lo que proporciona una visión más precisa de la distribución del ingreso en la región estudiada. Esta capacidad facilita la detección precisa de asentamientos informales y áreas vulnerables, lo que representa una herramienta muy útil para la sociedad civil en su conjunto al permitir la segmentación de zonas de bajos ingresos.

Las implicaciones de este trabajo son significativas en múltiples niveles. La capacidad de generalizar el modelo a otras áreas urbanas puede ayudar a cerrar la brecha de datos en zonas rurales y de bajos ingresos, incluyendo ciudades en países que no recolectan información mediante encuestas, como buena parte de los países de África subsahariana (Burke et al., 2021). Además, los mapas generados pueden facilitar la evaluación de impacto de políticas y programas localizados, así como también evaluar y mejorar la focalización de programas sociales. Disponibilizar una base de datos a nivel de grilla para el público general, como la que generé en este trabajo, facilita el acceso a datos geográficos sobre el ingreso, abriendo numerosas aplicaciones en diversas disciplinas más allá de la economía.

El desarollo de esta metodología presenta cuatro áreas clave para futuras investigaciones con el objetivo de mejorar estos modelos. En primer lugar, es crucial ampliar la cobertura geográfica del conjunto de datos, extendiéndolo más allá del AMBA. A medida que avanza la tecnología satelital y los costos asociados disminuyen, esta expansión se vuelve más viable en el futuro cercano.

En segundo lugar, la expansión temporal de las imágenes —-es decir, utilizar imágenes del AMBA para una serie de años en vez de un único punto en el tiempo— permitirá construir series de tiempo más largas y con mayor desagregación temporal y, además, las subsecuentes predicciones podrían mejorar su precisión ya que se podrían detectar cambios generados aleatoriamente por fallas en la imagen de un año en particular.

En tercer lugar, el modelo actual realiza predicciones de ingresos sin considerar la presencia de viviendas en las imágenes, lo que puede generar imprecisiones. Una posible solución a esto es la implementación de una red neuronal de dos etapas. En la primera etapa, la red identificaría la presencia de hogares dentro del área de interés, mientras que en la segunda etapa, si se detectan hogares, se procedería a predecir el ingreso per cápita. Esta estrategia no solo mejoraría la interpretación y la precisión de los resultados, sino que también permitiría calcular promedios por radio censal considerando únicamente celdas que contienen viviendas, reduciendo así el error general de la estimación.

Por último, explorar modelos más avanzados, como los *Vision Transformers* o redes convolucionales más potentes, podría mejorar aún más el rendimiento del modelo.

Puesta a disposición de las bases de datos generadas

Los mapas del ingreso generados para 2013, 2018 y 2020 se encuentran disponibles para su descarga y libre utilización en este repositorio.[12] A su vez, se disponibilizan los códigos para su replicación en este repositorio.[13] Debido a que las predicciones para cada celda de forma individual presentan cierta varianza entre predicción y predicción, se recomienda utilizar las estimaciones promediando las predicciones para áreas de interés (municipios, barrios, secciones o radios censales) y no a nivel individual. Como expuse a lo largo de este trabajo, los resultados agregados, incluso en áreas pequeñas como radios censales, predicen de forma precisa el ingreso de los hogares. De igual manera, en el repositorio es posible acceder y utilizar los parámetros entrenados del modelo, para generar predicciones sobre imágenes satelitales diferentes.

Referencias

Adamuz Peña, M. d. l. M. and González Tejeda, L. (2016), 'Demanda de vivienda de los hogares en méxico', *El trimestre económico* **83**(330), 311–337.

Blumenstock, J., Cadamuro, G. and On, R. (2015), 'Predicting poverty and wealth from mobile phone metadata', *Science* **350**(6264), 1073–1076.

Burke, M., Driscoll, A., Lobell, D. B. and Ermon, S. (2021), 'Using satellite imagery to understand and promote sustainable development', *Science* **371**(6535).

Center for International Earth Science Information Network (CIESIN) (2018), Documentation for the Gridded Population of the World, Version 4 (GPWv4), Revision 11 Data Sets, Technical report, Columbia University.

Chi, G., Fang, H., Chatterjee, S. and Blumenstock, J. E. (2022), 'Microestimates of wealth for all low- and middle-income countries', *Proceedings of the National Academy of Sciences* **119**(3), e2113658119.
URL: *https://www.pnas.org/doi/abs/10.1073/pnas.2113658119*

Ciaschi, M. (2021), 'Análisis distributivo utilizando información satelital. El caso de Argentina', *Estudios económicos* **38**(77), 5–38.

Dozat, T. (2016), 'INCORPORATING NESTEROV MOMENTUM INTO ADAM'.

Elbers, C., Lanjouw, J. O. and Lanjouw, P. (2003), 'Micro-Level Estimation of Poverty and Inequality', *Econometrica* **71**(1), 355–364.

Engstrom, R., Hersh, J. and Newhouse, D. (2022), 'Poverty from Space: Using High Resolution Satellite Imagery for Estimating Economic Well-being', *The World Bank Economic Review* **36**(2), 382–412.

Gasparini, L., Gluzmann, P. and Tornarolli, L. (2022), 'Caracterización de la población vulnerable: una propuesta con estimaciones para argentina', *Económica* **68**, 028.

Goodfellow, I., Bengio, Y. and Courville, A. (2016), *Deep Learning*, MIT Press. http://www.deeplearningbook.org.

Grosh, M. E. and Baker, J. L. (1995), *Proxy means tests for targeting social programs*, The World Bank.

Hansen, J. L., Formby, J. P. and Smith, W. J. (1998), 'Estimating the income elasticity of demand for housing: A comparison of traditional and lorenz-concentration curve methodologies', *Journal of Housing Economics* **7**(4), 328–342.

Henderson, J. V., Storeygard, A. and Weil, D. N. (2012), 'Measuring Economic Growth from Outer Space', *American Economic Review* **102**(2), 994–1028.
URL: *https://pubs.aeaweb.org/doi/10.1257/aer.102.2.994*

Huang, S., Tang, L., Hupy, J. P., Wang, Y. and Shao, G. (2021), 'A commentary review on the use of normalized difference vegetation index (ndvi) in the era of popular remote sensing'.

Instituto Nacional de Estadística y Censos (2003), 'La nueva Encuesta Permanente de Hogares de Argentina'.

Jean, N., Burke, M., Xie, M., Davis, W. M., Lobell, D. B. and Ermon, S. (2016), 'Combining satellite imagery and machine learning to predict poverty', *Science* **353**(6301), 790–794.

Khachiyan, A., Thomas, A., Zhou, H., Hanson, G., Cloninger, A., Rosing, T. and Khandelwal, A. K. (2022), 'Using Neural Networks to Predict Microspatial Economic Growth', *American Economic Review: Insights* **4**(4), 491–506. Publisher: American Economic Association.

LeCun, Y., Bengio, Y. and Hinton, G. (2015), 'Deep learning', *Nature* **521**, 436–444.

LeCun, Y., Bottou, L., Bengio, Y. and Haffner, P. (1998), 'Gradient-based learning applied to document recognition', *Proceedings of the IEEE* **86**, 2278–2324.
URL: *http://ieeexplore.ieee.org/document/726791/*

Pesaresi, M., Ehrlich, D., Ferri, S., Florczyk, A. J., Freire, S., Halkia, M., Julea, A., Kemper, T., Soille, P., Syrris, V. et al. (2016), *Operating procedure for the production of the Global Human Settlement Layer from Landsat data of the*

[12] https://doi.org/10.5281/zenodo.11200070

[13] https://github.com/Queeno11/Satellite-Imagery-Income-Prediction/releases/tag/v3.0.0

epochs 1975, 1990, 2000, and 2014, Publications Office of the European Union Luxembourg.

Piaggesi, S., Gauvin, L., Tizzoni, M., Cattuto, C., Adler, N., Verhulst, S. G., Young, A., Price, R., Ferres, L. and Panisson, A. (2019), Predicting city poverty using satellite imagery, *in* 'CVPR Workshops'.
URL: *https://api.semanticscholar.org/CorpusID:198182531*

Prince, S. J. D. (2024), 'Understanding deep learning'.
URL: *http://udlbook.com.*

Rao, J. and Molina, I. (2015), Empirical bayes (EB) method, *in* 'Small Area Estimation', John Wiley Sons, Ltd, chapter 9, pp. 269–332.

Rolf, E., Proctor, J., Carleton, T., Bolliger, I., Shankar, V., Ishihara, M., Recht, B. and Hsiang, S. (2021), 'A generalizable and accessible approach to machine learning with global satellite imagery', *Nature Communications* **12**(1). arXiv: 2010.08168 Publisher: Nature Research.

Rolnick, D., Veit, A., Belongie, S. and Shavit, N. (2017), 'Deep Learning is Robust to Massive Label Noise'. arXiv: 1705.10694.

Sheehan, E., Meng, C., Tan, M., Uzkent, B., Jean, N., Lobell, D., Burke, M. and Ermon, S. (2019), 'Predicting economic development using geolocated wikipedia articles'.

Shorten, C. and Khoshgoftaar, T. M. (2019), 'A survey on image data augmentation for deep learning', *Journal of Big Data* **6**.

Smythe, I. S. and Blumenstock, J. E. (2022), 'Geographic microtargeting of social assistance with high-resolution poverty maps', *Proceedings of the National Academy of Sciences* **119**(32).

Steele, J. E., Sundsøy, P. R., Pezzulo, C., Alegana, V. A., Bird, T. J., Blumenstock, J., Bjelland, J., Engø-Monsen, K., de Montjoye, Y.-A., Iqbal, A. M., Hadiuzzaman, K. N., Lu, X., Wetter, E., Tatem, A. J. and Bengtsson, L. (2017), 'Mapping poverty using mobile phone and satellite data', *Journal of The Royal Society Interface* **14**, 20160690.

Tan, M. and Le, Q. V. (2019), 'Efficientnet: Rethinking model scaling for convolutional neural networks'. arXiv: 1905.11946.

Tornarolli, L. (2018), Series comparables de indigencia y pobreza: una propuesta metodológica, Technical report, CEDLAS, Documento de Trabajo Nro. 226.

Ural, S., Hussain, E. and Shan, J. (2011), 'Building population mapping with aerial imagery and gis data', *International Journal of Applied Earth Observation and Geoinformation* **13**(6), 841–852.

Vrabel, J. (1996), 'Multispectral Imagery Band Sharpening Study', *Photogrammetric, Engineering and Remote Sensing* **66**(1), 73–79.

Weber, I., Kashyap, R. and Zagheni, E. (2018), Using advertising audience estimates to improve global development statistics, Technical report. Publication Title: ITU Journal: ICT Discoveries, Special Issue Issue: 2.
URL: *https://www.itu.int/en/journal/002/Pages/default.aspx*

Yarotsky, D. (2018), 'Universal approximations of invariant maps by neural networks'. arXiv: 1804.10306.

Yeh, C., Perez, A., Driscoll, A., Azzari, G., Tang, Z., Lobell, D., Ermon, S. and Burke, M. (2020), 'Using publicly available satellite imagery and deep learning to understand economic well-being in africa', *Nature Communications* **11**, 2583. ResNet -18 de imágenes de 6.72 x 6.72km de Landsat (224 x 224px) . Predicen Survey-measured asset wealth index y Census-measured asset wealth (PCA). Combinan day y nightime images.

APÉNDICE

A. *Selección de hiperparámetros*

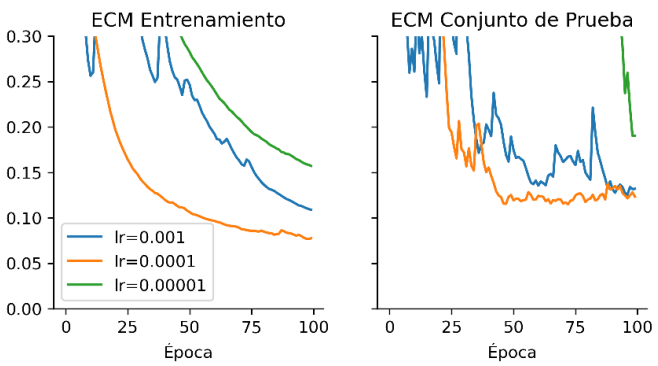
Para seleccionar el modelo con mejor rendimiento predictivo, se evaluaron una serie de combinaciones de hiperparámetros, buscando encontrar el conjunto que generaba mejores resultados. Sin embargo, dada la cantidad de combinaciones de hiperparámetros posibles, y el importante tiempo de cómputo que cada modelo implica (aproximadamente 50hs de entrenamiento), el conjunto de experimentos se limitó a una serie de pruebas concretas. Todos los experimentos se realizaron con imágenes de 100x100 metros (comprimidas a 128x128px) para reducir la carga computacional, y el modelo se entrenó por 100 épocas. Como base, los modelos utilizaron por defecto una tasa de aprendizaje de 0.0001, la arquitectura *EfficientNetV2S*, las cuatro bandas de color (RGB+NIR) e imágenes únicamente de 2013. En cada una de las pruebas, alguno de estos hiperparámetros se modifica, manteniendo los demás con esta configuración.

Los hiperparámetros a evaluar son los siguientes: tasa de aprendizaje (*learning rate* entre 0.001, 0.0001 o 0.00001), el tamaño de la red (*EfficientNetV2S*, *EfficientNetV2M* o *EfficientNetV2L*), la cantidad de bandas de color a utilizar (RGB o RGB+NIR) y el año de las imágenes utilizadas para el entrenamiento (solo 2013, 2013+2018 o 2013+2018+2022). Entonces, por ejemplo, para la primera prueba que analiza el impacto de la tasa de aprendizaje, se entrenan por 100 épocas tres modelos que varían la tasa entre 0.001, 0.0001 y 0.00001, la arquitectura *EfficientNetV2S*, las cuatro bandas de color (RGB+NIR) e imágenes únicamente de 2013.

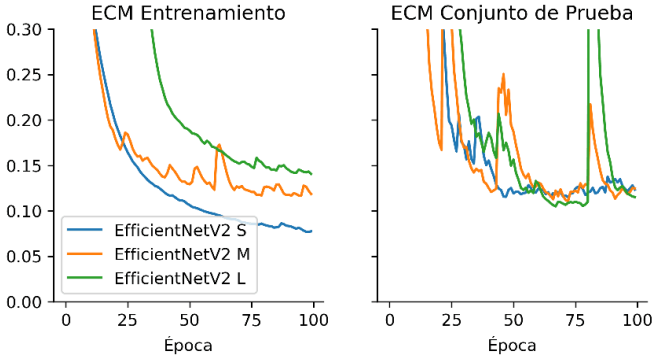
La Figura 15a muestra, para distintas tasas de aprendizaje, la evolución del error cuadrático medio a través de las épocas, para el conjunto de prueba. Para calcular el error del conjunto de prueba se calculó el promedio de las predicciones de cada radio censal y se comparó con el ingreso de cada radio censal, según la ecuación 6. El hiperparámetro seleccionado es una tasa de aprendizaje de 0.0001 por ser la que converge más rápido y presentar un menor MSE para prácticamente todas las épocas.

La Figura 15b replica la figura previa pero para diferentes tamaños de modelo. Como no se observa una clara mejoría en la evaluación de ninguno de las arquitecturas, se decidió mantener la más pequeña, ya que es la que requiere menor tiempo para su entrenamiento. La Figura 15c replica el mismo experimento cambiando las bandas de las imágenes. La inclusión de la banda cercana-infrarrojo parece reducir la varianza entre época y época de la performance del modelo, por lo que se decidió incluirla en el entrenamiento final del modelo.

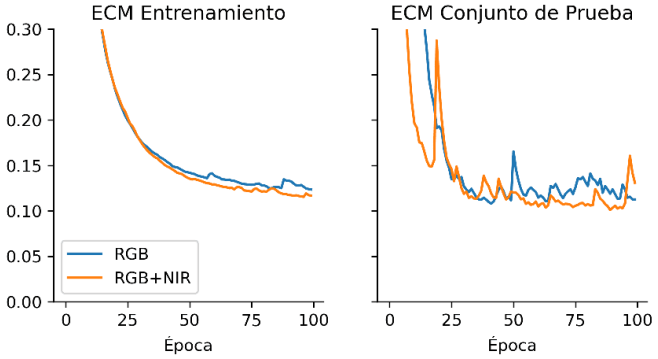
Finalmente, la Figura 15d compara el desempeño del modelo utilizando imágenes de diferentes años. El objetivo de incluir estas imágenes es que, si bien aportan un sesgo en la información presentada (ya que se mostrarán imágenes de, por ejemplo, 2022, con el ingreso de 2010), es posible que permitan una mejor generalización si se asume que no existen grandes diferencias en el contenido de las imágenes de los diferentes años; es decir, si se asume que los cambios urbanos llevan tiempo, y por lo tanto el sesgo de mostrar imágenes con 10 años de diferencia puede ser bajo. Es interesante que incluir imágenes "futuras" del AMBA permite obtener un mejor



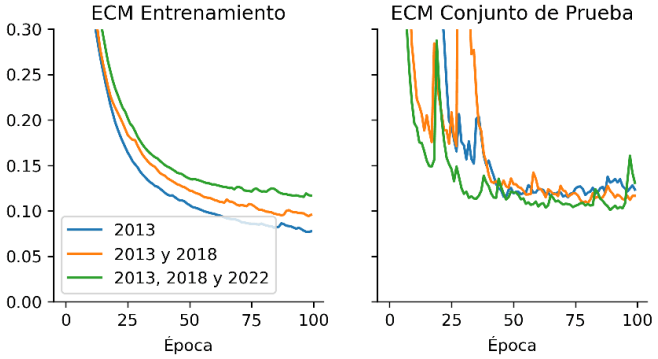
(a) Tasa de aprendizaje.



(b) Arquitectura del modelo.



(c) Bandas de las imágenes.



(d) Años de las imágenes.

Figura 15: Selección de hiperparámetros: Curvas de entrenamiento para los diferentes experimentos realizados.

desempeño del modelo, y una mayor consistencia a la hora de predecir en esos tres períodos. Si bien no se presenta en esta sección, si no se incluyen las imágenes de los tres años y se entrena solo con 2013, las posteriores predicciones para 2018 y 2022 resultan mucho más volátiles, posiblemente por tener diferencias de luz, ángulo de la captura satelital y vegetación que son interpretadas por el modelo como indicadores.

B. Tamaño de imágenes utilizadas

Una vez definidos los hiperparámetros, se buscó definir un tamaño de imagen óptimo para la predicción en forma de grilla. Si bien uno desearía obtener la predicción más pequeña y granular posible, es esperable que a medida que el área de predicción sea menor, la variabilidad en las predicciones pueda aumentar. De hecho, si se utilizan tamaños de imagen muy pequeñas (por ejemplo, predecir en un tamaño de 10x10mts) es probable que los resultados sean malos debido a que muchas propiedades exceden el tamaño de la imágen.

En este sentido, se evalúa el rendimiento de modelos con imágenes de 50x50mts, 100x100mts y 200x200mts, donde este último tamaño es muy similar al de muchos radios censales. El tamaño en pixeles de las imágenes siempre se mantiene en 128x128px para garantizar que los resultados sean comparables. Asimismo, se evalúa una estrategia que podría permitir resolver el problema de la escala: combinar imágenes de tamaño reducido (por ejemplo 50x50mts) con imágenes de mayor tamaño (por ejemplo 200x200mts), utilizando una mayor cantidad de bandas en las imágenes. En otras palabras, en vez de utilizar una imagen de 4 bandas, se construye una composición de 8 bandas donde las primeras 4 bandas se corresponden a la imagen de menor tamaño y las siguientes 4 a una de mayor tamaño. La Figura 16 ejemplifica este proceso con una imagen del AMBA.

Combinar información de áreas pequeñas con mayor resolución y áreas grandes con menor resolución permite que el modelo procese la información de ambas imágenes de forma conjunta, sin necesidad de modificar la arquitectura utilizada. Esto podría reducir la varianza de las predicciones, o bien mejorarla al considerar factores relevantes al ingreso que pueden no incluirse en la imagen muy cercana. Por ejemplo, es probable que si se comparan dos barrios con edificaciones similares, si uno se encuentra muy cerca de una autopista —típicamente ruidosa— se trate de un hogar de ingresos algo menores que en el caso alternativo.

La Tabla III compara el R^2 del conjunto de prueba para los diferentes tamaños de imágenes mencionados. Se utilizaron, en línea con lo descripto en la metodología, los parámetros de la época que minimiza el ECM en el conjunto de validación. Los seis modelos seleccionados tienen un poder predictivo alto, con todos superando valores de 0.84. El modelo de mejor performance, en línea con lo esperado, es el que combina las imágenes de menor tamaño ($50x50mts$) con las de mayor tamaño ($200x200mts$), que presenta un R^2 de $0{,}878$. Ese modelo es el utilizado en la sección de resultados. Sin embargo, es destacable que la mejoría relativamente pequeña, mejorando $0{,}03$ puntos el valor del indicador.

A pesar de esto, aunque es difícil de evaluar formalmente, al comparar los mapas generados con los diferentes tamaños de imágenes se observa una estimación más consistente al utilizar imágenes de $50x50mts$ combinadas con las de $200x200mts$. La Figura 17 muestra, a modo de ejemplo, el mapa de predicciones de 2013 para el modelo utilizando imágenes de $50x50mts$ a la izquierda, comparándolo con la combinación de $50x50mts$ y $200x200mts$ a la derecha. Debe prestarse especial atención a la correlación espacial entre las diferentes predicciones: mientras que en las predicciones con imágenes de $50x50mts$+$200x200mts$ se observan grupos consistentes de celdas pertenecientes al mismo decil, para las imágenes de $50x50mts$ existe una variabilidad mucho mayor en las predicciones, incluso cuando se encuentran en zonas aledañas. Si bien no existen datos para probar formalmente cuál es la distribución correcta, la variabilidad entre predicciones al utilizar imágenes tan pequeñas es esperable, ya que la información capturada por una sola imagen es muy reducida, ignorando el contexto de la misma.

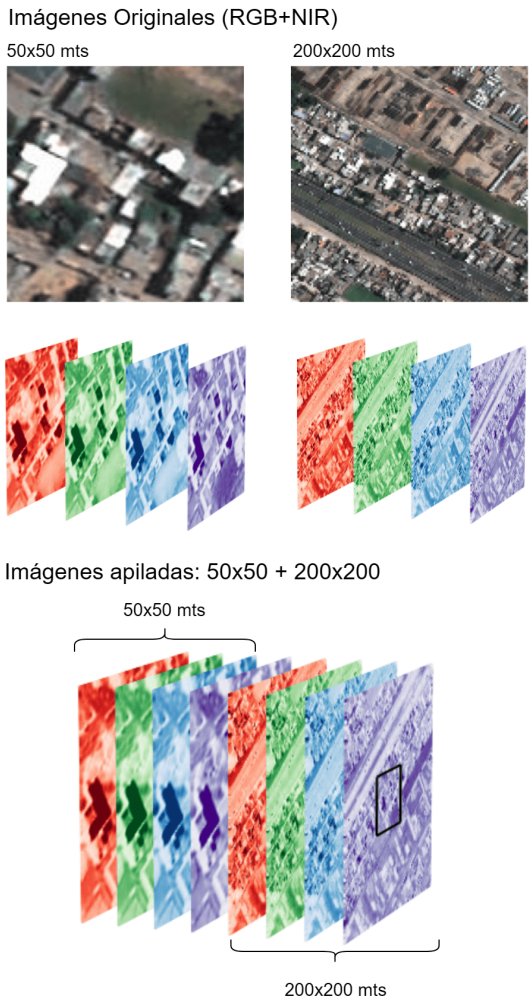
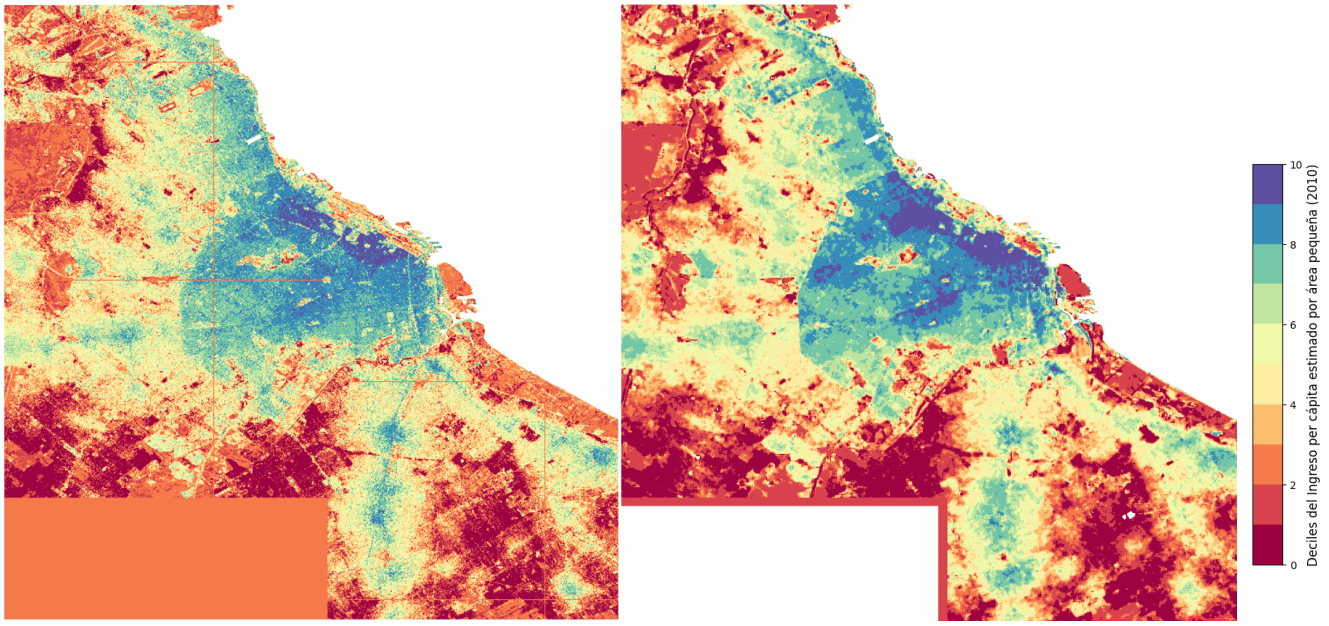


Figura 16: Construcción de imágenes apiladas.

Fuente: Imagen satelital provista por la CONAE.

C. Entrenamiento y Evaluación del modelo

La Figura 18 muestra la evolución de los errores cuadráticos medios a lo largo del entrenamiento, para el conjunto de en-



(a) 50×50 metros. (b) $50 \times 50 + 200 \times 200$ metros.

Figura 17: Estimación espacial del ingreso per cápita con diferentes tamaños de imagen.

Tabla III: Comparación de resultados para diferentes tamaños de imagen y configuraciones.

		Imagen apilada		
		Ninguna	100x100 mts	200x200 mts
Imagen base	50x50 mts	0.847	0.858	**0.878**
	100x100 mts	0.860	-	0.863
	200x200 mts	0.849	-	-

treamiento y de validación. La curva azul refleja la función de pérdida utilizada en el entrenamiento: el error cuadrática medio entre cada imagen y el ingreso del radio censal (Ecuación 4). La curva naranja presenta esta misma métrica calculada, pero calculada sobre el conjunto de validación. Finalmente, la curva verde es el error cuadrático medio promediando todas las predicciones de un radio censal sobre el conjunto de validación y comparando con el ingreso de ese radio (Ecuación 5). Como se mencionó previamente, el error calculado al nivel de radio censal es menor que el calculado para cada imagen, ya que pueden existir variaciones en el ingreso dentro del radio censal que, al promediarlas, dejan de impactar en la métrica. En estos casos, si el ingreso entre imágenes varía, como el ingreso per cápita promedio del radio censal es el mismo para todas las imágenes, el error será más alto. La sección IV-B discute en detalle la diferencia entre las dos métricas, y por qué la curva de validación por radio censal se encuentra sistemáticamente por debajo que las de validación y entrenamiento por imágen.

Los parámetros del modelo seleccionados para generar los resultados de este trabajo son los obtenidos en el *checkpoint* de la época 141 del entrenamiento, ya que se trató de la época que presentó un menor error cuadrático medio por radio censal sobre el conjunto de validación.

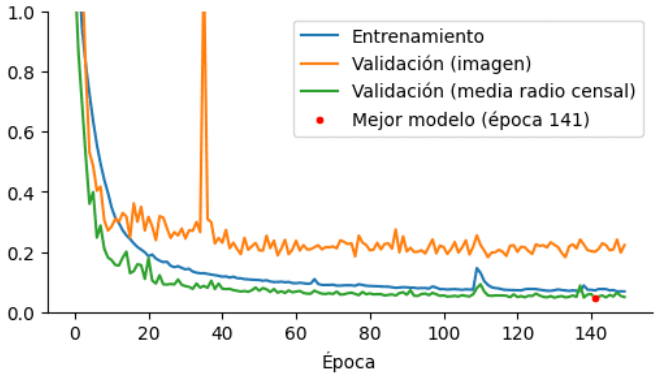


Figura 18: Error cuadrático medio a lo largo del entrenamiento.