UNIVERSIDAD
TORCUATO DI TELLA

Nowcasting de la pobreza en Argentina para el período 2016-2020

Tesis de Grado

Departamento de Economía

Licenciatura en Economía

Felipe Andersch

Juan Cruz Colombo

Francisco Granda

Wenceslao Sesma Lasierra

Tutora: Magdalena Cornejo

Índice

Índice de tablas

Índice de figuras

Resumen

En Argentina, la disponibilidad de microdatos provenientes de la Encuesta Permanente de Hogares (EPH) permite estimar las tasas de pobreza de la población urbana con un rezago de dos trimestres, lo que dificulta tanto la programación de las distintas políticas públicas de corto plazo como la evaluación de la efectividad de éstas. Es por ello que el objetivo de nuestra tesis es evaluar el desempeño de distintos modelos de series temporales que contemplen el uso de métodos de *Machine Learning* para realizar pronósticos en tiempo real (*nowcasts*) del ingreso per cápita por decil y, en forma indirecta, de la tasa de pobreza y pobreza extrema en la Argentina entre 2016 y 2020. En particular, se evaluará el desempeño de los pronósticos obtenidos a través de: *Bridge Equations* con selección de predictores a través de: (a) *Least Absolute Shrinkage and Selection Operator* (LASSO), (b) modelos de vectores autorregresivos aumentados por factores (FAVAR) usando el análisis de componentes principales para identificar dichos factores, (c) modelos dinámicos de factores utilizando el filtro de Kalman y (d) modelos autorregresivos integrados de medias móviles (ARIMA) que serán utilizados como *benchmark*.

Abstract

In Argentina, the availability of microdata from the Permanent Household Survey (EPH) makes it possible to estimate the poverty rates of the urban population with a lag of two quarters. Therefore, the planning and impact evaluation of short-term public policies, to reduce poverty, is not effective. That is why the objective of our thesis is to evaluate the performance of different time series models that contemplate the use of Machine Learning methods to make real-time forecasts (nowcasts) of per capita income per decile and, indirectly, of the rate of poverty and extreme poverty in Argentina between 2016 and 2020. In particular, the performance of the forecasts obtained through: Bridge Equations with selection of predictors using: (a) Least Absolute Shrinkage and Selection Operator (LASSO), (b) factor-augmented autoregressive vector models (FAVAR) using principal component analysis to identify such factors, (c) dynamic factor models using the Kalman filter, and (d) integrated moving average autoregressive models (ARIMA) that will be used as a benchmark.

1. Introducción

Nuestra principal motivación en esta tesis es monitorear de forma continua las tasas de pobreza y de pobreza extrema en Argentina, con el fin de contribuir a mejoras en el diseño de políticas públicas efectivas que permitan aliviar la situación de los más desfavorecidos en un contexto de condiciones económicas volátiles, en particular durante el período de la pandemia COVID-19.

En Argentina, la disponibilidad de microdatos provenientes de la Encuesta Permanente de Hogares (EPH) permite estimar las tasas de pobreza de la población urbana con un rezago de dos trimestres, lo que dificulta tanto la programación de las distintas políticas públicas de corto plazo como la evaluación de la efectividad de éstas.

El objetivo de nuestra tesis es evaluar el desempeño de distintos modelos de series temporales que contemplen el uso de métodos de *Machine Learning* para realizar pronósticos en tiempo real (*nowcasts*) del ingreso per cápita por decil y, en forma indirecta, de la tasa de pobreza y pobreza extrema en la Argentina entre 2016 y 2020. A partir de la construcción de de bases de datos de gran escala que consideran distintas variables macroeconómicas, publicadas en una frecuencia más alta que la EPH, como potenciales predictores se pronosticará el ingreso medio total familiar per cápita por decil a partir del cual se estimará la tasa de pobreza utilizando microsimulaciones.

En particular, se evaluará el desempeño de los pronósticos obtenidos a través de: *Bridge Equations* con selección de predictores a través de: (a) *Least Absolute Shrinkage and Selection Operator* (LASSO), (b) modelos de vectores autorregresivos aumentados por factores (FAVAR) usando el análisis de componentes principales para identificar dichos factores, (c) modelos dinámicos de factores utilizando el filtro de Kalman y (d) modelos autorregresivos integrados de medias móviles (ARIMA) que serán utilizados como *benchmark*.

El trabajo está organizado de la siguiente manera. En la sección 2 se realiza una revisión de la literatura acerca de las principales técnicas de *nowcasting* y trabajos empíricos más relevantes, con especial énfasis en mediciones de pobreza. En la sección 3 se describe la medición de la pobreza en la Argentina por los organismos oficiales, desde la recopilación de datos hasta sus procesamientos. En la sección 4 se desarrollan las cuatro técnicas utilizadas para realizar los ejercicios de *nowcast.* En la sección 5 se evalúa el desempeño de las distintas técnicas de pronóstico utilizadas. En la sección 6 se presentan las estimaciones de las tasas de pobreza a partir

de las proyecciones de ingresos por decil aplicadas sobre los microdatos. Por último, la sección 7 concluye.

2. Revisión de la literatura

En esta sección realizaremos una revisión sobre la literatura acerca de las principales técnicas de *nowcasting* y trabajos empíricos más relevantes, para luego hacer especial énfasis en la aplicación de dichas metodologías sobre las estimaciones y predicciones de la pobreza en tiempo real.

El trabajo pionero en desarrollar las técnicas de *nowcasting* dentro del ámbito de la economía es el de Banbura et al. (2013) en el que describe al *nowcasting* y su origen como la predicción del presente, el futuro cercano y el pasado más reciente. El término "*nowcasting*" surge como contracción de los términos "*now*" (ahora) y "*forecasting*" (pronóstico). Sin embargo, sus orígenes son anteriores a dicho trabajo y se dieron en el ámbito de la meteorología para luego ser adoptada en otras disciplinas como la economía.

Blanco et al. (2015) afirman que el principio básico del *nowcasting* es producir estimaciones tempranas, que se puedan actualizar secuencialmente, sobre variables de interés con frecuencias más altas a las de sus publicaciones. Esto se puede lograr a través de la utilización de la información contenida en una gran variedad de indicadores acerca del ciclo de una economía disponibles con frecuencias más altas que la variable a pronosticar.

Banbura et al. (2013) hacen énfasis en el objetivo y los obstáculos que los ejercicios de *nowcasting* deben sobreponerse. Al objetivo ellos lo explican como el ejercicio de leer a través de un filtro generado por un modelo, un flujo de información en tiempo real. Con esto se busca formalizar factores importantes acerca de cómo los agentes económicos leen, interpretan y reaccionan ante datos en tiempo real. Los *nowcasts* deben poder manejar problemas de dimensionalidad y distintas frecuencias de datos, dada la vasta cantidad de información y las frecuencias con que los distintos predictores están disponibles.

De acuerdo con Banbura et al. (2013) existen dos categorías de ejercicios de *nowcasting* en economía: modelos conjuntos de representación del espacio de estado y modelos parciales.

Los modelos conjuntos de representación del espacio de estado tienen como idea principal que hay factores no observables que impulsan a las variables dependientes, es decir las variables que nos interesan estimar. Este modelo trata a cada serie temporal como la suma de dos componentes ortogonales: el primero, impulsado por un conjunto de factores no observados captura la dinámica conjunta y el segundo es tratado como un residuo idiosincrático. Esta metodología implica que el *nowcast* puede realizarse siguiendo un proceso en dos etapas. En el primer paso, se estiman factores comunes para los datos disponibles. En el segundo paso, se utilizan dichos factores como regresores de la variable de interés. Este tipo de modelos permite rastrear fuentes de revisiones de los pronósticos. Incluso, en caso de ser necesario, ante la publicación simultánea de información de variables explicativas podemos descomponer y cuantificar la contribución de cada noticia individual o un conjunto de ellas sobre la actualización del *forecast*.

Con respecto a la utilización de modelos conjuntos de representación del espacio en *nowcasts*, el objetivo de este es construir un marco para la lectura del flujo de publicaciones de datos en tiempo real. Dentro de este contexto el modelo produce *forecasts* para todas las variables en las que estamos interesados, de manera que permite extraer nueva información a partir de los componentes no observados de los nuevos datos publicados. De esta manera el modelo provee un método clave para entender los cambios en las estimaciones de la actividad económica en tiempo real y contribuye a evaluar cuán significativo es cada componente observado.

Los modelos parciales, por su parte, utilizan el método más simple y antiguo de *nowcasting*, según Banbura et al. (2013). Se caracterizan por no especificar un modelo conjunto para las variables predictoras y las de interés, también indican la necesidad de estos modelos de ser complementados con modelos auxiliares o de un conjunto de parámetros separados para cada serie de datos. Es por esto, que una limitación importante de este tipo de modelos es que, dado que no hay una representación en conjunto de las variables, no es posible cuantificar el impacto de datos/noticias específicas sobre el *nowcast.* De todos modos, a pesar de las limitaciones mencionadas los modelos parciales tienen una larga tradición en instituciones políticas como lo son los bancos centrales. Dentro de estos modelos encontramos *Bridge Equations* y *MIDAS*, mediante los cuales los *nowcasts* de la variable de interés son obtenidos utilizando regresiones, que en el caso de *Bridge Equations* el predictor es una variable agregada de menor frecuencia y

para paliar los problemas de distintas frecuencias deben ser utilizados modelos del estilo ARMA o VAR. Respecto a los modelos *MIDAS,* los predictores son incluidos en la regresión con la frecuencia observada (y, en general, en frecuencia más alta que la variable a pronosticar) y a través de polinomios de rezagos asigna un peso empírico sobre las observaciones de los predictores en alta frecuencia.

La literatura acerca de *nowcasting* ha permitido afirmar tres conclusiones generales. En primer lugar, los *nowcasts* mejoran progresivamente su precisión a medida que se acerca la fecha de publicación de la información acerca de la variable de interés. En segundo lugar, la explotación de determinados datos que resultan relevantes para el análisis permite mejorar la exactitud de los *nowcasts*. Por último, el uso de procedimientos automatizados mostró tener desempeños tan buenos como las estimaciones institucionales.

El trabajo de Giannone et al. (2008) es pionero en el desarrollo del método de *nowcasting* para datos macroeconómicos. Los autores desarrollan un modelo de pronósticos con la finalidad de abordar los inconvenientes que surgen por del uso de series de datos muy amplias que, además, son publicadas con diferentes frecuencias y retrasos. En su trabajo combinan la idea de "*bridging"* de la información mensual con el *nowcast* del producto bruto trimestral y utilizan una gran cantidad de datos disponibles. El problema de dimensionalidad y de grados de libertad lo resuelven explotando la colinealidad de los factores y concentrando la información en unos pocos factores. Los pronósticos que obtienen mediante este proceso permiten calcular y actualizar el *nowcast* del producto bruto trimestral cada vez que la información mensual se actualiza, además con la publicación de cada dato actualizado es posible calcular el efecto marginal de esa nueva información publicada sobre el pronóstico del *nowcast* y su precisión.

Un ejemplo de *nowcast* realizado para variables macroeconómicas de Argentina es el trabajo publicado por Blanco et al. (2015), en él explican que como consecuencia de que las cifras del PIB cuentan con un retraso significativo en su publicación, con el uso de distintas técnicas de *nowcasting* se puede obtener una percepción inmediata de la tendencia del ciclo económico. Para esto utilizan dos enfoques: ecuaciones puente (*bridge equations*) y un modelo multivariado de factores, que luego comparan, para concluir que el *nowcast* basado en un modelo de factores supera en predictibilidad al de *bridge equations*.

En años más recientes, las técnicas de *nowcasting* empezaron a aplicarse para monitorear y estimar la tasa de pobreza. En este sentido, Aguilar et al. (2019) y Aiken et al. (2020) explican que la utilidad de la aplicación de dichas metodologías sobre las estimaciones y predicciones de la pobreza en tiempo real radica en que cálculos oportunos y comparables acerca de los niveles de pobreza cumplen un rol vital a la hora de poder evaluar el nivel de desarrollo de un país, y permitir a los distintos organismos una asignación de sus recursos más eficiente para así poder cumplir sus objetivos. La tarea de determinar quién es elegible y quién no para distintos planes sociales y ayuda humanitaria es una ineficiencia importante a la hora de administrarlos. Generalmente estos programas sociales toman las decisiones acerca de a quienes asignarles la ayuda a través de registros impositivos e información específica acerca de cada individuo de la población que se pueda encontrar en organismos de un país, suelen ser poco confiables y no recientes.

En esta línea, los siguientes trabajos de investigación que citamos son una pequeña muestra de distintos ejercicios de *nowcast* de la tasa de pobreza utilizando métodos de *Machine Learning.* Algunas de estas técnicas las utilizaremos en nuestro trabajo de investigación para pronosticar el ingreso per cápita familiar por decil complementando los modelos tradicionales de series temporales y aprovechando distintas variedades de información disponibles.

Aiken et al. (2020) tienen como objetivo principal responder a la siguiente pregunta: ¿pueden los datos no tradicionales ser usados para mejorar la eficiencia de ayuda social? Mediante el uso de datos no tradicionales de celulares y otros sensores digitales, demuestran que métodos de aprendizaje supervisado son tan precisos como cualquier estimación basada en encuestas estándar a la hora de identificar pobreza, consumo y niveles de riqueza.

En cambio, Luccheti (2018) utiliza métodos supervisados de *Machine Learning* para estimar la dinámica de la riqueza de los hogares ante la ausencia de datos de panel. La metodología propuesta consiste en estimar parámetros de un modelo de ingresos logarítmicos en la primera ronda de datos transversales usando Lasso, para predecir los ingresos de hogares encuestados en la segunda ronda pero que falta su información de la primera ronda. El autor sugiere que utilizar un procedimiento de regularización como Lasso ante la ausencia de datos longitudinales acerca de individuos o hogares en dos o más momentos del tiempo genera resultados alentadores para estimar la movilidad económica.

En un trabajo reciente, Brum y De Rosa (2021) calculan las transferencias necesarias para neutralizar el efecto de la crisis sobre la tasa de pobreza. Proponen realizar estimaciones micro simulando los efectos de corto plazo de la crisis sobre la tasa de pobreza actual y los efectos de las políticas públicas sobre ella. Un aspecto positivo de este enfoque es que provee una consistencia micro-macro entre impactos heterogéneos sobre los hogares y el shock al PBI, además de hacer posible calcular los efectos específicos a trabajadores informales y/o autónomos.

Por su parte, Sampi y Jooste (2020), proponen un indicador distinto e innovador, el "Google Mobility Index", para pronosticar tasas de crecimiento de producción industrial mensual en economías de América Latina y el Caribe. En el trabajo se utiliza un "backcasting" como metodología para incrementar el número histórico de observaciones y luego aumenta el rezago a una semana en el dato de movilidad en conjunto con otros datos de alta frecuencia. Finalmente, la regresión con una muestra de datos mixtos se implementa para pronósticos de tasas de crecimiento de la producción industrial. Resultando el índice de movilidad de Google en un buen predictor de la producción industrial.

En sintonía con el objetivo de este trabajo, Parolin y Wimer (2020) realizan un pronóstico de pobreza durante la crisis del Covid-19 para Estados Unidos. Para esto realizan un pronóstico de las tasas de pobreza utilizando dos definiciones diferentes de pobreza, ambas usando un marco de *Suplemetary Poverty Measure* (SPM). El primero es la medida de pobreza estándar de SPM, que incorpora todos los impuestos y transferencias. El segundo es antes de impuestos, medida previa a la transferencia de la pobreza de la SPM utilizando los umbrales de pobreza contemporáneos. SPM es una metodología que presenta sustanciales mejoras con respecto a la medida oficial de pobreza vigente en Estados Unidos desde 1960.

Desde otra perspectiva Silalahi (2020), realiza un pronóstico de datos de pobreza utilizando modelos ARIMA estacionales en la provincia de Java Occidental. En este trabajo SARIMA es el desarrollo del modelo ARIMA añadiendo un efecto estacional. ARIMA se considera más popular en la predicción porque es más flexible y según la autora representa muchas variaciones de datos en una serie de tiempo particular. Mientras que SARIMA es el desarrollo del modelo ARIMA que tiene un efecto estacional. Esto surge como solución ya que el patrón de los datos de pobreza cada año tiene un efecto estacional según las condiciones económicas y los cambios de posición en el

gobierno, por lo que puede afectar la percepción del gobierno sobre las políticas implementadas para la reducción de la pobreza.

Por último, el estudio quizás más disruptivo acerca de la medición de la pobreza en Argentina es el publicado por Cardinale Lagomarsino et al. (2016) en el que a través del algoritmo de aprendizaje *Random Forest* aplicado a la Encuesta Permanente de Hogares (EPH) logran predecir con un 85% de efectividad la condición de pobreza en un hogar. Para lograr estos resultados incorporan metodologías de robustez y realizan correcciones al desbalanceo muestral de su variable de interés, el ingreso.

3. Datos

Las nociones de pobreza e indigencia empleadas para la medición de la tasa de pobreza y de pobreza extrema, respectivamente, en Argentina son presentadas mediante el enfoque de la Línea de Pobreza y Línea de Indigencia. Este enfoque metodológico corresponde al método de medición indirecta o "línea" determinando a partir del costo monetario de una canasta de alimentos capaz de cubrir el umbral mínimo de necesidades alimenticias (Canasta Básica Alimenticia), siendo aquellos hogares que no lo superen considerados hogares en pobreza extrema. De la misma manera, para calcular la línea de pobreza se extiende el umbral para incluir consumos básicos no alimentarios, así determinando la Canasta Básica Total (CBT), construida en base a los hábitos de consumo alimentarios y no alimentarios (vestimenta, transporte, educación, salud, etcétera) de la población de referencia. Dado que la CBA y CBT se construyen por hogar, el valor de éstas es luego contrastado con los ingresos totales de cada hogar para determinar las tasas de pobreza y de pobreza extrema.

Entonces, para poder estimar en tiempo real la tasa de pobreza y pobreza extrema es importante tener pronósticos del ingreso medio per cápita familiar el cual será comparado con las líneas de pobreza e indigencia para determinar si una persona es pobre o está dentro de la pobreza extrema. Es a través de la Encuesta Permanente de Hogares (EPH) realizada por el INDEC que se construye el ingreso por adulto.

La EPH tiene como objetivo caracterizar a la población en términos de sus características sociodemográficas, inserción en la producción social de bienes, servicios y participación en la distribución del producto social. Intenta lograr esto a través de encuestar una pequeña fracción representativa del total de los hogares, a nivel nacional tiene una extensión de 25.000 hogares por trimestre y 100.000 hogares por año, en la que se renueva periódicamente el conjunto de hogares a encuestar y en la que la aplicación de técnicas estadísticas permite garantizar la precisión de los datos obtenidos. Cada encuesta sucede trimestralmente y brinda información para la ventana de observación, el trimestre, teniendo así una modalidad de relevamiento continuo. El esquema elegido para la EPH continua se lo ha llamado 2-2-2, dado que su funcionamiento es el siguiente:

- Las viviendas en un área ingresan a la muestra para ser encuestados en dos trimestres consecutivos, en el mes y semana asignados a esa área.
- Se retiran por dos trimestres consecutivos.
- Vuelven a la muestra para ser encuestados en dos trimestres consecutivos en el mes y semana asignados a esa área.[1]

Es importante mencionar que, si bien los datos oficiales permiten estimar la pobreza en forma trimestral, el INDEC usa y recomienda trabajar con datos bimestrales para evitar la estacionalidad (explicada en gran parte por el medio aguinaldo) y para asegurar una adecuada precisión a nivel de aglomeración.

3.1. Medición de la pobreza y la pobreza extrema en Argentina

El INDEC siempre realizó mediciones oficiales acerca de la pobreza y pobreza extrema, sin embargo, no se puede establecer la evolución de estos indicadores a lo largo del tiempo porque no presentan series históricas completas o comparables entre sí. La falta de comparabilidad se debe a razones como:

1. Modificaciones en la EPH, como en el año 2003 que se modificaron aspectos del muestreo, temáticos de los cuestionarios, y organizativas.

[1] La nueva Encuesta Permanente de Hogares de Argentina.2003 - INDEC

2. Modificaciones en la metodología de medición de la pobreza y pobreza extrema, como en el año 2016 que se modificaron los patrones de consumo con los que se eligen los alimentos que integran la CBA y CBT.
3. Durante el período 2007-2015 no se utilizó información fiable.

Es por eso por lo que Gasparini et al. (2019) realizan el esfuerzo de presentar series comparables que permiten documentar las tendencias que ha seguido la pobreza en Argentina durante las últimas décadas a través de las mediciones habituales de pobreza oficial por ingreso y otras mediciones alternativas, obteniendo resultados desalentadores, el nivel de pobreza de ingresos al año 2018 es superior al existente en 1983.

En el primer enfoque, la pobreza de ingresos concluye que la movilidad de ingresos en el corto plazo es relevante por la mayor volatilidad macroeconómica y menor protección de instituciones laborales y oportunidades de acceso al crédito. Además, quienes se encuentran en los deciles más bajos de ingresos son quienes sufren mayor oscilación de sus ingresos, es por eso que cambios marginales en sus salarios reales puede ubicar una parte significativa de la población como pobreza.

En el segundo enfoque mide la pobreza multidimensional a través de la metodología propuesta por Santos y Villatoro (2018), que puede ser aplicado a pesar de las limitaciones que presenta la EPH, ya que esta no fue diseñada para monitorear la pobreza multidimensional. Esta metodología consiste en medir las privaciones/carencias de los hogares respecto a 5 dimensiones: características habitacionales, acceso de servicios básicos de infraestructura, acceso a educación, empleo y protección social, e ingresos. Concluye que el porcentaje de la población que se encontraba en situación de pobreza, bajo este enfoque, para los años entre 2003 y 2018 se redujo sustancialmente, aunque la reducción fue a una tasa de cambio mayor en los primeros años. Resulta importante notar que una característica propia de este tipo de pobreza es que suelen ser impasibles a los vaivenes macroeconómicos porque tienen en cuenta las dimensiones estructurales de la pobreza.

En el tercer enfoque estima la pobreza crónica, caracterizada por carencias que no pueden ser superadas aún bajo condiciones económicas coyunturalmente favorables persistentes, a través de una metodología que asocia pobreza crónica con alta vulnerabilidad, dado las limitaciones de la EPH. Los hogares con características tales que es improbable que eviten la situación de pobreza

de ingresos son considerados de alta vulnerabilidad. Se busca tener una mirada extendida en el tiempo de la pobreza de ingresos. Quienes se encuentran en situación de pobreza crónica suelen ser cuentapropistas no calificados y asalariados en pequeñas firmas que trabajan en la construcción, servicio doméstico y comercio.

Respecto a la medición de la pobreza en Argentina y el desarrollo de distintas metodologías con el objetivo de lograr una mejor estimación de los parámetros de interés, encontramos diversas fuentes que proveen esta información aparte del INDEC. Grandes esfuerzos han sido realizados por instituciones como el CEDLAS en colaboración con el Banco Mundial, el Observatorio de la Deuda Social de la Universidad Católica Argentina, y más recientemente Martín González Rozada comenzó a difundir a través de las redes sociales estimaciones en tiempo real de la tasa de pobreza.

El CEDLAS le encuentra dos problemas a la metodología utilizada por el INDEC: la identificación y la agregación de los pobres.

En la guía SEDLAC (base de datos socioeconómicos para América Latina y Caribe), observan que a la hora de la identificación de los pobres no existen claros argumentos normativos u objetivos para fijar una línea debajo de la cual todos son pobres y sobre la cual todos son no pobres. Existe una arbitrariedad fundamental en la definición de pobreza, que diferentes autores y agencias usan, por lo tanto, proponen la utilización de un rango de líneas. Especifican que las mediciones de pobreza adoptada por los países difieren en el criterio utilizado para identificar a los pobres, y que las mediciones internacionales son los instrumentos inevitables para comparar niveles de pobreza absolutos y tendencias en países, regiones y a nivel mundial.

Respecto a la agregación, para cada línea de pobreza computan los tres indicadores de pobreza más usados: la tasa de pobreza, la brecha de pobreza y el Índice de pobreza Foster-Greer-Thorbecke, calculando la pobreza en base a la distribución de los individuos y no de los hogares.

Por otro lado, el Observatorio de la Deuda Social Argentina de la UCA describe en su documento estadístico: Pobreza Monetaria y Vulnerabilidad de Derechos, la medición de la pobreza a través de la Encuesta de la Deuda Social Argentina (ODSA) la cual surge de un diseño muestral probabilístico de tipo polietápico estratificado y con selección sistemática de viviendas, hogares y población en cada punto de muestra (5760 hogares). Resaltan además que los índices elaborados por el ODSA son diferentes respecto a los elaborados por INDEC y metodológicamente

independientes. Algunos puntos importantes donde difieren significativamente son: el diseño muestral, la estrategia ante la no repuesta del hogar, tamaño muestral, área de cobertura, estrategia del relevamiento.

En años más recientes, Martín González Rozada es quien ha logrado reconstruir la serie más larga de pobreza para nuestro país, desde el año 1974 hasta la actualidad. Para esto reconstruye las series de ingresos por individuo, y utiliza diferentes valores de canastas según las variaciones metodológicas y de medición de la EPH o mediciones anteriores similares según corresponda a los datos relevados para cada momento del tiempo. En su metodología para la medición del ingreso, González-Rozada y Menéndez (2002), comienzan por considerar el ingreso familiar per cápita como la suma de los ingresos por trabajador del grupo familiar y el ingreso no laboral del mismo. El ingreso laboral en este caso es modelado a través del ingreso laboral individual de cada uno de sus miembros y en particular trabajan con la distribución del ingreso como función de la participación, el desempleo, la educación y retornos sobre las características individuales. Además, los autores utilizan micro simulaciones con el objetivo de evaluar contrafácticos y otros efectos en la distribución general de la función del ingreso.

3.2. El ingreso medio per cápita familiar por decil

Los datos de ingreso per cápita familiar fueron obtenidos a partir de los microdatos de las EPH del INDEC. Posteriormente fueron calculados los valores medios por decil y se tomaron promedios semestrales para evitar la estacionalidad. Una vez calculados estos valores tomamos la tasa de crecimiento logarítmica trimestral a partir de la cual se desarrollarán los distintos modelos de pronóstico.

Es importante destacar que los datos correspondientes al tercer trimestre del 2007 no están disponibles en las bases de datos del INDEC debido a causas de orden administrativo, por este motivo procedimos a promediar la tasa de crecimiento anual de los últimos 3 años anteriores a esa fecha para cada decil y aplicarla al dato del 2006T3, y de esta forma conseguir los datos faltantes del 2007T3.

Para afrontar el problema de la falta de datos de la EPH entre 2015T3 y 2016T1 optamos por hacer una interpolación de estos a través de un modelo ARIMA seleccionado en forma automática en R. El algoritmo se explica y comenta posteriormente en este documento.

3.3. Potenciales predictores del ingreso

Para determinar las tasas de pobreza en la Argentina, el INDEC utiliza la metodología de líneas de indigencia (LI) y líneas de pobreza (LP), es decir, determina una línea a través de la cual aquellos individuos que no la sobrepasen son considerados indigentes o pobres. Para calcular la línea de indigencia se utiliza la Canasta Básica Alimentaria (CBA), que en cierta medida representa el umbral mínimo de necesidades energéticas y proteicas para subsistir. En caso de no contar con los ingresos suficientes para poder consumir la totalidad de esta canasta, se considera perteneciente al grupo de indigentes. A su vez, la línea de pobreza no solo se concentra en los alimentos necesarios para la vida diaria, sino que también toma en cuenta otros consumos básicos no alimentarios. La suma de ambos compone la Canasta Básica Total (CBT), en caso de no tener ingresos suficientes para comprar esta canasta, se pertenece al grupo de pobres. La CBA y la CBT son informadas mensualmente. Los bienes y servicios que componen estas canastas son revalorizados utilizando los datos del Índice de Precios al Consumidor (IPC). El ingreso de los individuos u hogares es capturado en la Encuesta Permanente de Hogares (EPH), pero como mencionamos anteriormente, este dato se pública con rezagos. Es por ello, que utilizaremos variables macroeconómicas para producir una predicción del ingreso por decil.

Las variables macroeconómicas que utilizamos son 24. Cada una fue elegida en función de si podía tener un efecto sobre los ingresos teniendo en cuenta la teoría económica. Los datos de las variables elegidas abarcan el periodo que va desde el primer trimestre del 2004 hasta el último trimestre del 2020. Este grupo de variables se compone sobre datos de la producción, como el PIB y el EMAE; datos del consumo tanto público y privado; datos del empleo, del salario mínimo y planes sociales como la AUH; datos monetarios y financieros publicados por el BCRA, como la base monetaria, la tasa de plazos fijos y las reservas internacionales; datos sobre comportamientos de los consumidores como los depósitos y los préstamos tomados; y por último datos del tipo de cambio con respecto al dólar (oficial y “blue”).

A las variables seleccionadas se les aplicó la transformación necesaria para volverlas estacionarias. Como podemos ver en el Apéndice A, se utilizaron los tests de raíz unitaria ADF, PP y KPSS y a aquellas variables no estacionarias, se les aplicó una transformación de diferencias logarítmicas, de esta forma eliminando los efectos de tendencia. Por último, las variables elegidas tienen frecuencias de publicación trimestrales, mensuales o diarias. En caso de no tener frecuencia

trimestral, dependiendo de si la variable es de tipo stock o flujo, detallado en el Apéndice B, se les tomó la suma o el promedio simple, respectivamente. De esta forma llevando todas las variables a una frecuencia trimestral.

Una lista detallada con las fuentes de todas las variables se encuentra en la Tabla 1.

Tabla 1. Descripción de los Predictores Macroeconómicos

Dato	Frecuencia	Fuente
Producto Bruto Interno (millones de pesos, a precios de 2004)	Trimestral	Dirección Nacional de Cuentas Nacionales - INDEC
Estimador Mensual de Actividad Económica (Base 2004)	Mensual	Dirección Nacional de Cuentas Nacionales - INDEC
Consumo Nacional Privado (millones de pesos, a precios corrientes)	Trimestral	Dirección Nacional de Cuentas Nacionales - INDEC
Consumo Público (millones de pesos, a precios corrientes)	Trimestral	Dirección Nacional de Cuentas Nacionales - INDEC
Exportaciones FOB (millones de pesos, a precios corrientes)	Trimestral	Dirección Nacional de Cuentas Nacionales - INDEC
Indice Confianza del Consumidor (UTDT)	Mensual	CIF - UTDT
Tasa de Actividad (Total Aglomerados)	Trimestral	Encuesta Permanente de Hogares, INDEC
Tasa de Empleo (Total Aglomerados)	Trimestral	Encuesta Permanente de Hogares, INDEC
Tasa de Desocupación (Total Aglomerados)	Trimestral	Encuesta Permanente de Hogares, INDEC
Salario Mínimo, Vital y Móvil (en pesos corrientes)	Mensual	MTEySS
Asignación Universal por Hijo (en pesos corrientes)	Mensual	Ministerio de Hacienda sobre la base de ANSES y normas de actualización
Evolución del Empleo	Mensual	EIL-Ministerio de Trabajo
Tasa de Plazo Fijo 30 días (en % nominal anual)	Mensual	BCRA
Base Monetaria (millones de pesos)	Mensual	BCRA
Depósitos en Pesos (millones de pesos)	Mensual	BCRA
Depósitos en Dólares (millones de USD)	Mensual	BCRA
Préstamos al Sector Privado en Pesos (millones de pesos)	Mensual	BCRA
Préstamos al Sector Privado en Dólares (millones de USD)	Mensual	BCRA
Reservas Internacionales del BCRA (millones de USD)	Mensual	BCRA
Gasto Público Nacional en Política de Ingresos (en millones de pesos corrientes)	Mensual	Dirección de Análisis de Política Fiscal y de Ingresos. Subsecretaría de Programación Macroeconómica. Secretaría de Política Económica, Ministerio de Economía
Tipo de Cambio Paralelo (USD)	Diario	https://www.ambito.com/contenidos/dolar-informal-historico.html
Tipo de Cambio Oficial (USD)	Diario	https://www.ambito.com/contenidos/dolar-oficial-historico.html
Remuneración Promedio de los Trabajadores del Sector Privado (desestacionalizada)	Mensual	Fuente: Observatorio de Empleo y Dinámica Empresarial - MTEySS - en base a SIPA
Precio Soja Futuro EEUU (USD)	Mensual	https://es.investing.com/commodities/us-soybeans-historical-data

4. *Nowcasting* del ingreso medio per cápita familiar por decil

En esta sección se presentan distintas estrategias de *nowcasting* de la tasa de crecimiento trimestral del ingreso medio semestral per cápita familiar por decil que se utilizarán para luego evaluar cuál es el enfoque que mejor desempeño tiene en términos de pronóstico para el período comprendido entre 2016 y 2020.

El ejercicio de pronóstico en tiempo real (*nowcasting*) que proponemos intenta evaluar la capacidad predictiva del ingreso total familiar de los hogares (por decil) a través de distintas técnicas de *machine learning* que permiten reducir la dimensión de la información disponible junto con modelos tradicionales de series temporales.

Utilizaremos como *benchmark*, modelos univariados (i.e. modelos ARIMA) tradicionales de las series a pronosticar y los pondremos a competir con otros modelos que permitan incorporar otros predictores macroeconómicos medidos en alta frecuencia (mensual).

En cada uno de los enfoques se realizarán pronósticos recursivos para $h = 1$ pasos adelante (1 trimestre) usando como ventana inicial el período comprendido entre el primer trimestre de 2004 y el cuarto trimestre de 2015 ($T = 48$ observaciones). Por lo tanto, obtendremos 19 pronósticos correspondientes a los 19 trimestres comprendidos entre 2016Q1 y 2020Q3.

4.1. Modelos ARIMA (*benchmark*)

Los modelos *ARMA (*p, q) son modelos que combinan dos procesos estocásticos diferentes. Las siglas *AR* representan a procesos estocásticos auto-regresivos asociados con el parámetro *p*; mientras que las siglas *MA* representan procesos de promedios móviles asociados con el parámetro *q*. Estos modelos son representados por la ecuación:

$$y_t = a_0 + \sum_{i=1}^{p} a_i y_{t-i} + \sum_{i=0}^{q} \beta_i \varepsilon_{t-i} \qquad (1)$$

Si la serie temporal aplicada a la ecuación tiene raíces con valor absoluto mayor a uno, entonces se lo llama modelo *ARMA* para y_t . De todas maneras, si alguna de las raíces tiene valor absoluto menor o igual a uno, se dice que la secuencia y_t es un proceso integrado, y se llama a la ecuación (1) *ARIMA (*p, q, d), donde la I hace referencia a integrado. Los modelos ARIMA son

tales que diferenciados d vece se convierten en modelos *ARMA (*p, q). Este último paso tiene el objetivo de eliminar la no estacionariedad de la media.

Por otro lado, los modelos *SARIMA* son una extensión de los modelos *ARIMA*, que permiten modelar directamente el componente estacional de la serie de datos.

Para seleccionar el modelo *ARIMA* de mejor ajuste para la tasa de crecimiento medio del ingreso per cápita por decil utilizamos el algoritmo de Hyndman y Khandakar (2008) que combina pruebas de raíz unitaria, la minimización del criterio de Akaike corregido (AICc) y la estimación por máxima verosimilitud.[2] Este algoritmo sigue los siguientes pasos para determinar el modelo de mejor ajuste:

i. Se determina el número de diferenciaciones $0 \leq d \leq 2$ necesarias para volver estacionaria a la serie en función de la prueba KPSS de estacionariedad.

ii. Los valores p y q se eligen en función de la minimización del AICc luego de diferenciar d veces la serie temporal (según lo indicado en el paso anterior). En lugar de utilizar todas las combinaciones posibles de p y q (que son potencialmente infinitas), el algoritmo hace un step-wise search:

   a. Estima cuatro modelos iniciales:

      - $ARIMA(0, d, 2)$
      - $ARIMA(2, d, 2)$
      - $ARIMA(1, d, 0)$
      - $ARIMA(0, d, 1)$

      Se incluye una constante salvo que $d = 2$. Si $d \leq 1$, se adiciona el siguiente modelo inicial: $ARIMA(0, d, 0)$ sin constante.

   b. El mejor modelo (con el menor AICc) estimado en (a) se considerará el modelo actual.

   c. Adicionalmente, se consideran variantes del modelo actual:

      - variando p y q respecto del modelo actual en ± 1
      - incluyendo/excluyendo la constante del modelo actual.

[2] El algoritmo fue implementado en R usando la función auto.arima() de la librería "forecast".

El mejor modelo (ya sea el actual o alguna variante) se convierte en el nuevo modelo actual.

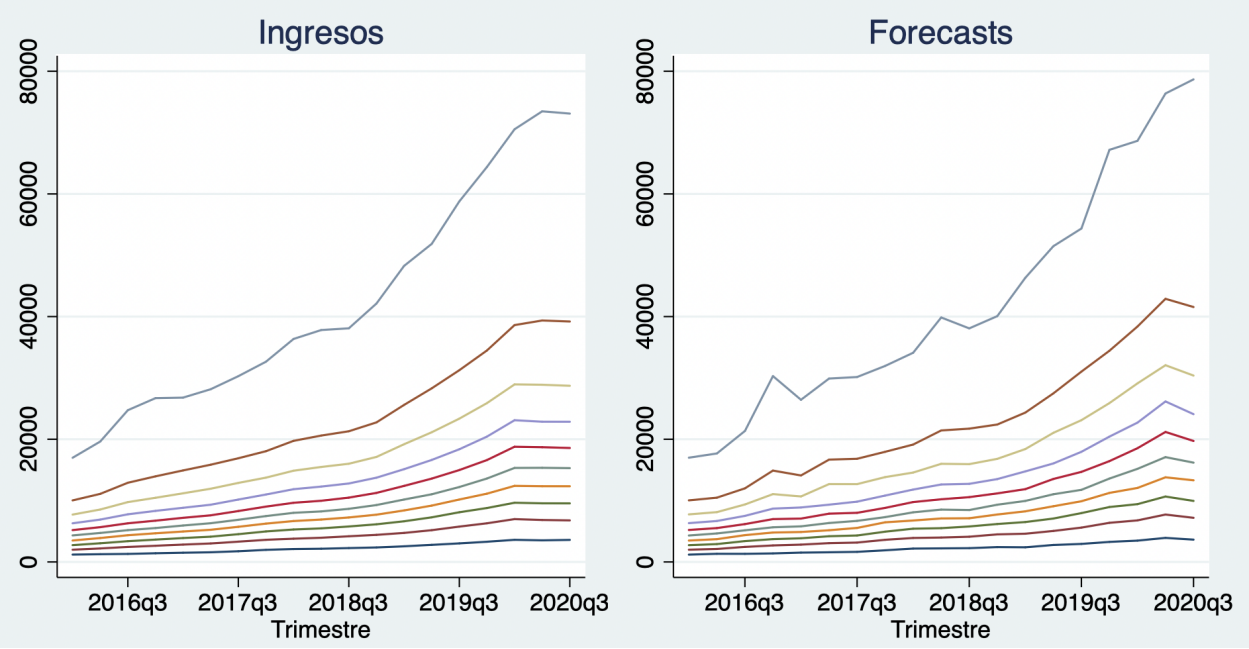
d. Se repite (c) tantas veces como sea necesario hasta que no se logre encontrar un menor AICc.

En nuestro caso, el algoritmo será modificado de forma tal que evalúe la posibilidad de modelar también el comportamiento estacional a través de los modelos *SARIMA*.

La Figura 1 muestra los pronósticos recursivos para $h = 1$ de las series de ingreso medio per cápita familiar por decil (tanto en nivel como en diferencias) durante el período de pronóstico (2016Q1 a 2020Q3) y los compara con los valores observados o actuales (Actual).

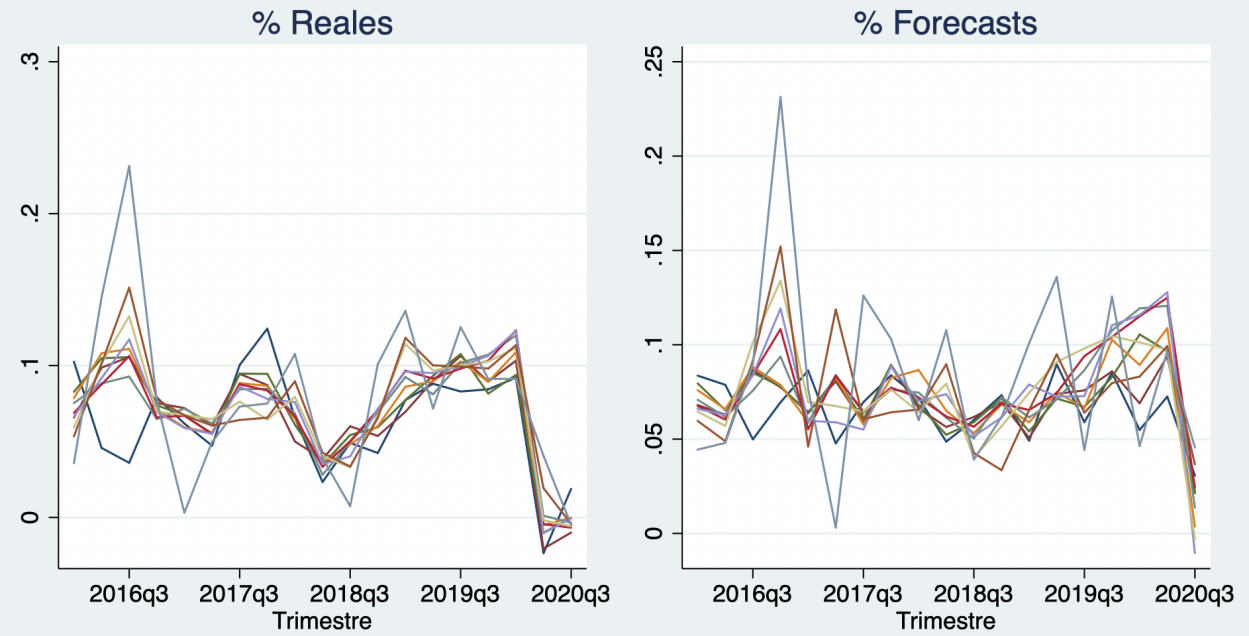
Figura 1. Pronósticos *ARIMA* recursivos para $h = 1$ (en niveles y diferencias logarítmicas)

(a) En niveles



Nota: en rojo se presentan los pronósticos y en negro los valores observados por decil.

(b) En diferencias logarítmicas



4.2. *Bridge equations* y selección por LASSO

Los modelos basados en *Bridge equations* son modelos dinámicos que explican la variable de baja frecuencia (en nuestro caso, el ingreso medio semestral por decil en frecuencia trimestral) a través de rezagos de una variable predictiva medida en frecuencia más alta (e.g. mensual). El *nowcast* de la variable de interés es obtenido a través de la siguiente regresión:

$$y_{t,1}^{k_1} = \propto + \beta y_{t,n}^{k_1} + e_t^{k_1}, \quad t = k_1, 2k_1, \ldots \qquad (2)$$

La variable $y_{t,1}^{k_1}$ es la de interés (i.e. el ingreso), $y_{t,n}^{k_1}$ es un predictor en alta frecuencia agregado en la frecuencia más baja de la variable de interés y $k_1$es la frecuencia con la que se publica la información de la variable de interés. Como se puede observar, el problema de tener frecuencias mixtas se soluciona a través de la agregación temporal de los predictores en alta frecuencia a la frecuencia más baja correspondiente a la variable de interés en pronosticar.

Para que las series de alta frecuencia coincidan con la frecuencia de la variable objetivo, sumamos los stocks, y promediamos los flujos (ver Apéndice B). Optamos por agregar los datos diarios y mensuales a la frecuencia trimestral utilizando un promedio aritmético.

Para poner este modelo en práctica, primero definimos la variable predictiva, agregada a la frecuencia más baja de la variable de interés como: $y_{t,n}^{k_1} = \sum_{i=1}^{j} \quad w_i x_t^{M_i}$, donde w_i representa la ponderación asociada a la publicación M_i. En segundo lugar, la ecuación (2) es estimada a través de la agregación trimestral de la información mensual a través de MCO (mínimos cuadrados ordinarios) ya que las variables explicativas están predeterminadas al estar publicadas con antelación a la variable a pronosticar. En caso de ser necesario este tipo de modelos se puede extender para incluir más predictores o rezagos de la variable dependiente y variable de interés (ver e.g. Mitchell, 2009).

En caso de tener muchos potenciales predictores de la variable de interés, como en esta tesis, se puede utilizar algún algoritmo de selección de variables como LASSO.

LASSO es un método introducido por Tibshirani (1996) que estima los coeficientes de un modelo lineal intentando reducir la dimensionalidad y regularizar el modelo a través de penalizaciones para prevenir *overfitting* (sobreajuste) mejorando así la capacidad predictiva del modelo. Entonces, LASSO logra mejorar la capacidad predictiva eliminando variables y reteniendo solo aquellas que logran mejorar el ajuste del modelo. Esta técnica, si bien puede inducir algún tipo de sesgo, puede bajar dramáticamente la varianza, así mejorando el ECM (error cuadrático medio). El objetivo de Lasso es resolver el siguiente problema de minimización de la suma de los residuos al cuadrado (como el estimador de MCO), pero sujeto a una función de penalización (una restricción sobre la normal L_1 del vector de coeficientes:

$$T^{-1} \sum_{t=0}^{T-1} \quad (y_{t+1} - \beta^T x_t)^2 + \lambda \sum_{i=1}^{K} \quad |\beta_i| \qquad (3)$$

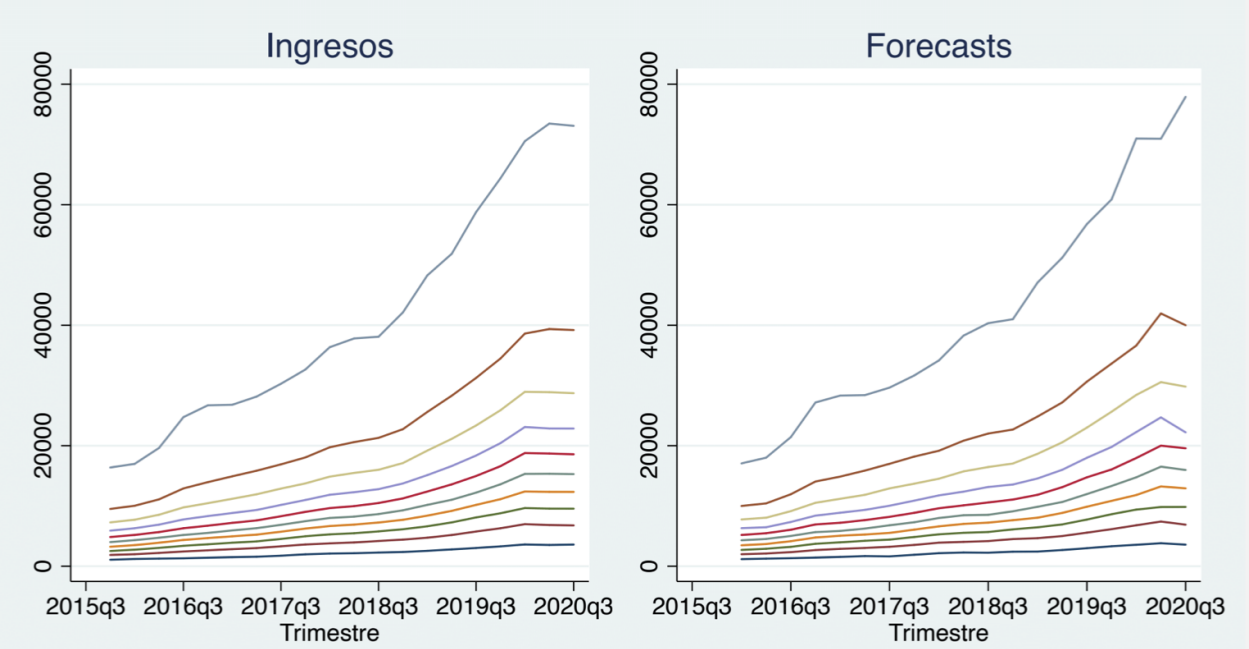
donde $\lambda \geq 0$ es un parámetro de complejidad que captura el peso (relativo) de la función de penalización y tiene el efecto de contraer los parámetros estimados hacia cero y *K* es el número total de predictores que se incluyen. Como todos los parámetros serán penalizados en la misma magnitud (ya que λ es un escalar) se asume que todos los predictores tienen varianzas que fueron

escaladas a 1. A mayor λ, mayor es la penalización ya que más coeficientes serán iguales a cero (excluidos del modelo) y más parsimonioso será nuestro modelo de predicción.

En la práctica, necesitamos seleccionar λ. En nuestro caso, el parámetro λ será seleccionado a través de un ejercicio de validación buscando aquel valor que logre minimizar el criterio de información de Akaike corregido (AICc), siguiendo la misma lógica aplicada para el modelo *benchmark*.[3]

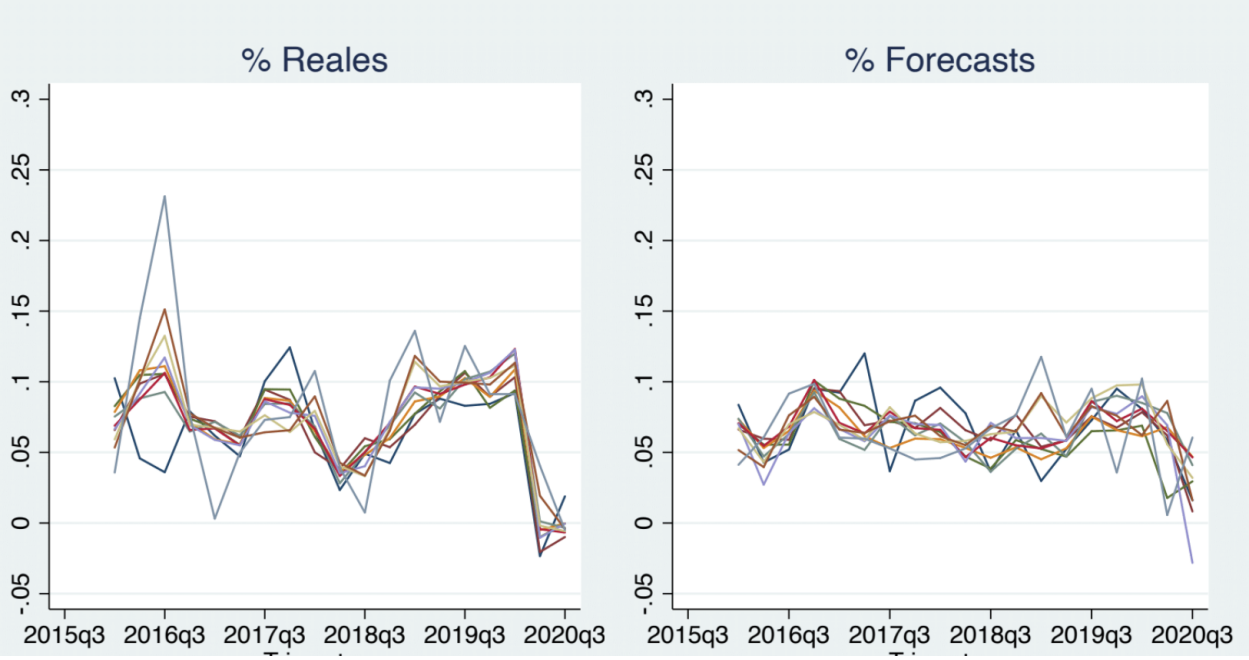
Figura 2. Pronósticos *Bridge Equations* recursivos para $h = 1$ (en niveles y diferencias logarítmicas)

(a) En niveles



[3] LASSO fue implementado en STATA 16 usando el comando `lasso2`.

(b) En diferencias logarítmicas



4.3. Modelos FAVAR usando el análisis de componentes principales

Nuestro objetivo principal es pasar del modelo predictivo (4) al modelo con factores (5), lograr una representación parsimoniosa que a la vez reduzca la dimensión resumiendo las fuentes de covariación entre las variables originales.

$$y_t = \sum_{i=1}^{N} \beta_i x_{it} + \lambda(L) y_t + \varepsilon_t, \quad 1, \dots, T \quad (4) \qquad y_t$$

$$= \sum_{s=1}^{q} \delta_s F_{st} + \lambda(L) y_t + \varepsilon_t \qquad 1, \dots, T \quad (5)$$

Notar que con (5) se logra crear una cantidad $q < N$ de factores, $F_{st} = X\delta$, y los coeficientes δ_s representan la ponderación de cada x_s sobre el factor F_s, e $\lambda(L)$ representa un vector de polinomios de rezagos. Es importante mencionar que a fines predictivos no importa la interpretación de los factores.

El análisis de componentes principales es una técnica multivariada que nos permite lograr esto resumiendo la gran cantidad de predictores que tenemos en unos pocos factores que captan la

máxima variación informacional de éstas, a través de la descomposición en autovalores-autovectores de sus matrices de correlaciones o de varianzas y covarianzas. Estos *s* factores, $F_{st} = X\delta$, son combinaciones lineales de los N vectores de predictores disponibles (*X*), que mejor reproduce la varianza de las variables originales del modelo predictivo (4). El primer factor, F_{1t} , está dado por una combinación de las variables originales que mejor captura su máxima varianza posible sujeto a la restricción de que los "pesos" o auto vectores tengan módulo igual a 1. El segundo factor, F_{2t} , captura la mayoría de la información no incluida en el primer factor, y a la vez no está correlacionado con este, son ortogonales entre sí. En este trabajo se seleccionan todos los factores que tengan autovalores mayores a uno porque esos factores explican la varianza total más que el promedio.[4]

Por otro lado, en los modelos FAVAR (6) se busca estimar modelos VAR, con la diferencia de que utiliza factores estimados por algún método de reducción de dimensionalidad, en nuestro caso, análisis de componentes de factores. Sean F_t el vector de los factores seleccionados, a través del método de análisis de componentes principales y G_t el vector de factores no observables del ingreso que son indicadores de conceptos intangibles.

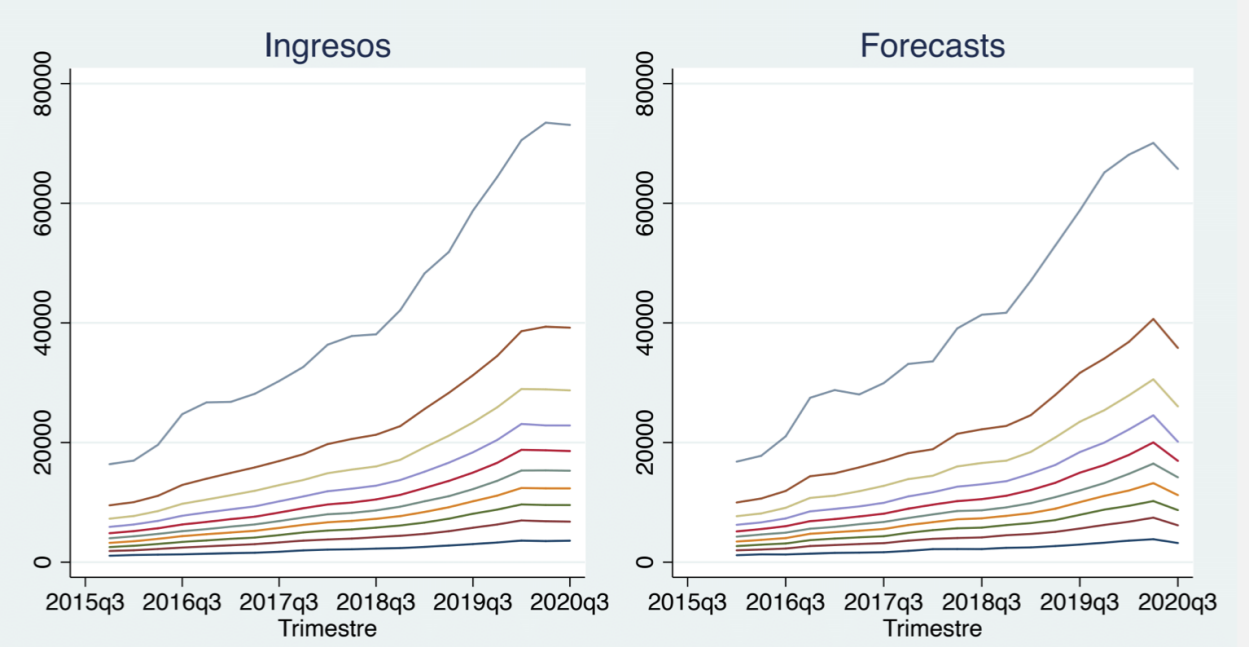
$$(F_t\ G_t) = \phi_1(F_{t-1}\ G_{t-1}) + \cdots + \phi_p\left(F_{t-p}\ G_{t-p}\right) + v_t \qquad (6)$$

Notar que ϕ_j es una matriz de coeficientes correspondiente a los términos autorregresivos de orden j. En nuestro caso, elegiremos el orden p del modelo VAR de acuerdo con el criterio de información de AIC.

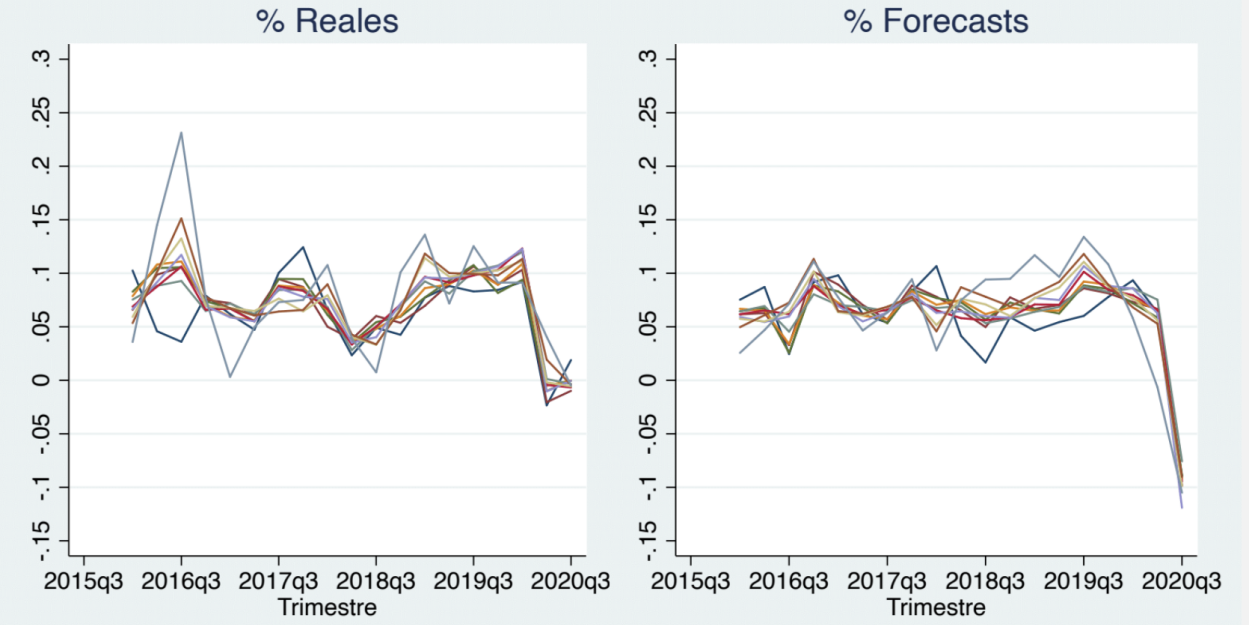
[4]Los Componentes Principales fueron calculados en STATA 16 usando el comando `pca`.

Figura 3. Pronósticos FAVAR recursivos para $h = 1$ (en niveles y diferencias logarítmicas)

(a) En niveles



(b) En diferencias logarítmicas



4.4. Modelos dinámicos de factores usando el filtro de Kalman

Este tipo de modelos se caracteriza por tener un modelo conjunto especificado para la variable de interés, Y_t (i.e. el ingreso), que tiene una representación de estado-espacio.

$$Y_t = \mu + \zeta_t(\theta)X_t + G_t, \qquad G_t \backsim i.i.d. \quad N\left(0, \sum_{G} \quad (\theta)\right) \qquad (7)$$

$$X_t = \phi_t(\theta)X_{t-1} + H_t, \qquad H_t \sim i.i.d. \quad N\left(0, \sum_{H} \quad (\theta)\right) \qquad (8)$$

La ecuación (7) vincula el vector de la variable observada, Y_t , con un vector de variables de estado no observadas, X_t, y la ecuación de transición (8) especifica la dinámica de estas últimas. Notar que las matrices de coeficientes, $\zeta_t(\theta)$ y ϕ_t, así como también las matrices de covarianzas de las perturbaciones, $\sum_H \quad (\theta)$ e $\sum_G \quad (\theta)$, varían con el tiempo.

Dadas las ecuaciones (7) y (8) que representan el modelo y el parámetro θ, el rol del filtro de Kalman es proveernos de las esperanzas condicionales de las variables no observables de estado, sujeto a el conjunto de información disponible Ω_Y, es decir:

$$X_{(\Omega_Y)} = E_\theta\{\Omega_Y\} \qquad (10)$$

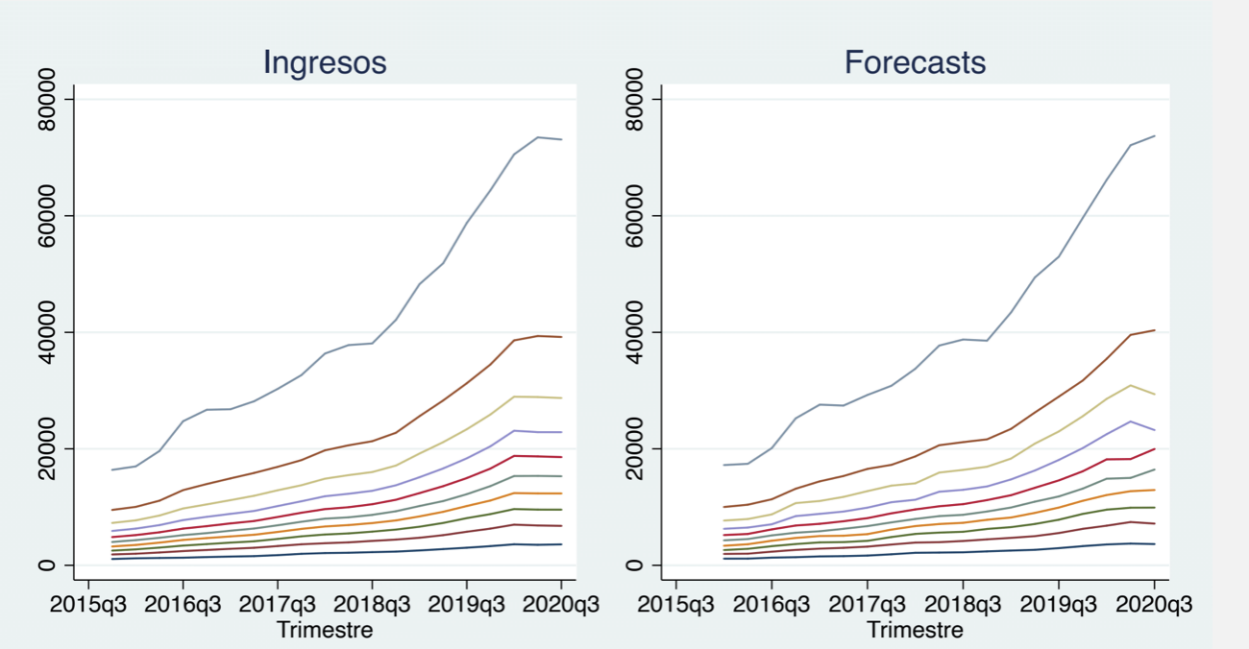
Es importante mencionar que el filtro de Kalman trabaja eficientemente con observaciones faltantes de Y_t, y nos provee de las esperanzas condicionales de esas observaciones faltantes.

Utilizaremos la metodología de dos pasos propuesta por Giannone et al. (2008) para estimar las ecuaciones (7) y (8). En primer lugar, los parámetros de la ecuación de estado-espacio son estimados usando el método de componentes principales, mencionado en la sección 4.3, realizado para una muestra "balanceada" de Y_t, considerando solo la parte de las muestras para el cual las observaciones de todas las variables de estado están completas. El segundo paso consiste en re-estimar los factores aplicando el filtro de Kalman para todo el conjunto de información disponible.[5]

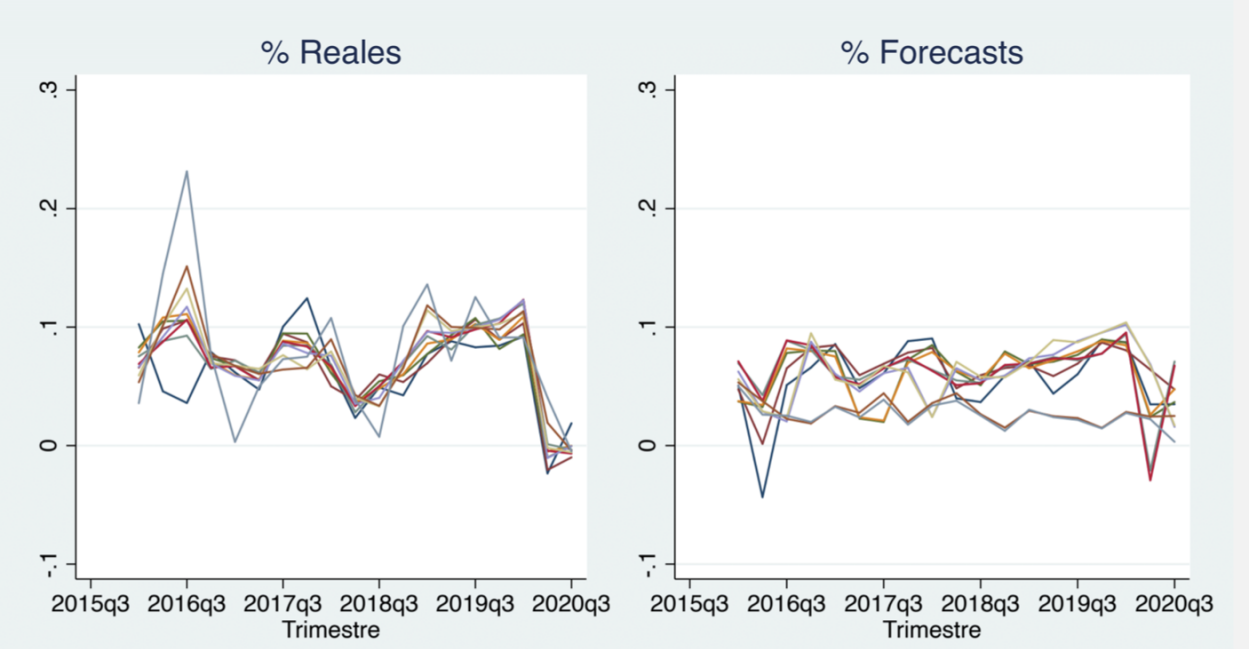
[5] El MDF fue implementado en STATA 16 usando el comando `dfactor`.

Figura 4. Pronósticos MDF recursivos para $h = 1$ (en niveles y diferencias logarítmicas)

(a) En niveles



(b) En diferencias logarítmicas



5. Evaluación del desempeño de los distintos pronósticos

La precisión de los pronósticos solo se puede determinar considerando qué tan bien se desempeña un modelo con datos nuevos que no se usaron al ajustar el modelo. Una práctica muy difundida en este campo es la de separar la base de datos disponible en lo que se conoce como datos de entrenamiento y test. Donde los datos de entrenamiento se usan para estimar los parámetros de un método de pronóstico y los datos de prueba o test se usan para evaluar su precisión.

Siguiendo las recomendaciones sobre qué medias deben ser utilizadas al momento de comparar la precisión de los modelos utilizados, las principales medidas de desempeño de pronósticos utilizadas en este documento son *Root Mean Square Error* (RMSE) y *Mean Absolute Percentage Error* (MAPE).

Con respecto a RMSE definido como RMSE = $\sqrt{mean(e_t^2)}$ es una de las medidas comúnmente utilizadas, en la cual su escala depende de la escala de los datos. En este caso resulta preciso utilizar esta herramienta ya que es muy útil para comparar entre diferentes métodos aplicados sobre el mismo set de datos. El RMSE es ampliamente reconocido por su relevancia teórica en el modelaje estadístico. De todos modos, es importante remarcar que la medida es más sensible a valores extremos que otras medidas como MSE (*mean square error*) y MdAE (*median absolute error*)

Por otro lado, el MAPE, medida basada en errores porcentuales y dado por MAPE = mean ($|P_t|$), tiene la ventaja de ser independiente de la escala y es frecuentemente utilizada para comparar la performance entre diferentes sets de datos. Es importante notar que esta medida tiene la desventaja de ser infinita o indefinida si $y = 0$ para cualquier t en el período de interés, y tener una distribución extremadamente sesgada cuando cualquier valor de y cercano a 0. También tienen la desventaja de imponer una penalización mayor a los errores positivos que a los negativos.

La evaluación de los modelos a través de la comparativa de los distintos errores medios por decil nos permitió llegar a la conclusión de que no hay un modelo que predomine sobre los otros. El modelo de mejor performance varía de acuerdo a cada decil y dependiendo de si tomamos el MAPE o el RMSE como parámetros de evaluación. Para profundizar el análisis y llegar a un mejor

resultado final nos dispusimos armar pools de modelos. El pool A es el resultado de tomar un promedio simple entre los modelos LASSO y KALMAN; para armar el pool B se agregó al promedio el pronóstico del modelo ARIMA y para el pool C se añadió el modelo MDF. Una vez creados los pools procedimos a evaluarlos mediante el RMSE y el MAPE, aunque el resultado no nos condujo a un pool de modelos que predomine sobre el resto.

Con el objetivo mejorar la precisión de nuestros pronósticos y, teniendo en cuenta que no pudimos concluir que un modelo prevalezca sobre el resto, nos dispusimos a crear un *pool* de modelos siguiendo un criterio de minimización de errores cuadráticos (MCO). A través de una minimización no lineal buscamos para cada decil las ponderaciones óptimas de cada modelo que nos lleven a minimizar el error cuadrático medio. Como restricciones impusimos la necesidad de que la suma de las ponderaciones sea igual a 1, y no permitimos que tomen valores negativos. El ejercicio nos condujo a una matriz de ponderadores óptimos para cada decil. Luego, evaluamos la precisión del nuevo modelo con el RMSE y el MAPE, obteniendo la mejor performance para cada decil teniendo en cuenta únicamente el RMSE. A la hora de evaluar el resultado con el MAPE obtuvimos la mejor performance con el modelo MCO para todos los deciles exceptuando el número 7 y número 9.

Tabla 2. Evaluación de Métodos

DECIL		MDF	KALMAN	LASSO	ARIMA	POOL A	POOL B	POOL C	MCO
1	RMSE	125,602281	73,254167	94,098899	113,485794	79,7339693	87,795022	92,4312826	**72,5937041**
	MAPE	3,27%	2,5920%	2,99%	2,97%	2,71%	2,5925%	2,64%	**2,49%**
2	RMSE	219,061181	190,730363	169,077941	243,455798	176,002117	195,308047	191,513752	**165,787729**
	MAPE	2,91%	2,68%	2,570%	2,73%	2,61%	2,59%	2,574%	**2,43%**
3	RMSE	279,185228	179,285995	182,038171	293,740715	170,541364	196,466162	196,366419	**160,161036**
	MAPE	2,82%	2,68%	2,53%	2,41%	2,44%	**2,31%**	2,32%	2,33%
4	RMSE	370,41	241,17	333,16	429,48	277,21	314,29	331,66	**210,52**
	MAPE	2,74%	2,71%	2,90%	2,78%	2,69%	2,48%	2,49%	**2,44%**
5	RMSE	421,501687	339,067312	382,090473	479,08405	307,785234	337,552753	366,531782	**215,808285**
	MAPE	2,46%	2,08%	2,55%	2,32%	2,18%	2,16%	2,30%	**1,76%**
6	RMSE	551,06307	415,53057	486,1833	663,9485	397,163586	441,621982	481,086498	**254,524486**
	MAPE	2,41%	2,11%	2,51%	2,82%	2,15%	2,34%	2,46%	**1,68%**
7	RMSE	798,820829	543,025742	577,373367	852,036476	539,050819	615,697724	616,082521	**527,522868**
	MAPE	2,82%	2,79%	2,5567%	2,73%	2,57%	2,5555%	**2,47%**	2,69%
8	RMSE	846,728081	654,969346	565,645419	907,67485	600,26451	677,256627	664,063787	**518,510036**
	MAPE	2,90%	2,89%	2,17%	3,09%	2,49%	2,56%	2,45%	**2,04%**
9	RMSE	1066,17323	1459,36655	933,771106	1152,52465	1103,77007	1003,65169	961,767662	**800,880575**
	MAPE	2,86%	4,80%	**2,59%**	3,48%	3,54%	3,23%	3,01%	2,77%
10	RMSE	2482,31268	2911,71343	2011,1483	2510,27429	2247,57671	2037,25311	1928,07534	**1548,29857**
	MAPE	4,39%	5,78%	3,93%	4,98%	4,68%	4,24%	4,07%	**3,65%**

* p<.10, ** p<.05, *** p<.01

Una vez evaluada la *performance* de nuestro modelo nos propusimos evaluar la significatividad de nuestros resultados. El test de Diebold-Mariano consiste en realizar una regresión por MCO de la diferencia entre el error cuadrático del modelo de mejor performance (MCO) y el modelo de segunda mejor performance para cada decil. La regresión se realiza únicamente contra una constante, siendo la significatividad de esta el parámetro por el cual podemos definir o no si es que nuestro modelo tiene el menor RMSE. Si la constante es significativa entonces podemos decir que existe una diferencia entre el RMSE del mejor modelo y del segundo mejor, por lo que estadísticamente hay evidencia para asegurar que el modelo MCO es el que presenta el menor error cuadrático medio. Por otro lado, si la constante es no significativa entonces no hay evidencia estadística que nos permita afirmar que la diferencia de los errores cuadráticos entre el modelo MCO y el modelo de segunda mejor performance sea distinta de cero.

El test fue llevado a cabo en STATA para cada decil, tomando siempre el modelo MCO y el modelo de segunda mejor performance, que varía dependiendo el decil (ver Tabla 2). Los resultados no nos permiten concluir que hay evidencia estadística de que el modelo de MCO muestre mejores resultados que el modelo de segunda mejor performance. La intuición nos lleva a pensar que el motivo por el cual no logramos obtener resultados significativos estadísticamente es la poca cantidad de observaciones que utilizamos, lo que resulta en un test de muy baja potencia.

6. Estimación de la pobreza en tiempo real

Esta sección de nuestro trabajo no ha sido finalizada aún. Esperamos poder terminarla durante el transcurso de las próximas 2 semanas.

6.1. Microsimulaciones

6.2. Caracterización de la pobreza entre 2016 y 2020

7. Conclusiones

Esta sección de nuestro trabajo no ha sido finalizada aún. Esperamos poder terminarla durante el transcurso de las próximas 2 semanas.

Referencias

Aguilar, R. A. C., Mahler, D. G., y Newhouse, D. (2019). Nowcasting Global Poverty. IARIW-World Bank Working Paper.

Aiken, E., Bedoya, G., Coville, A., y Blumenstock, J. E. (2020). *Targeting Development Aid with Machine Learning and Mobile Phone Data*. Working Paper. University of California, Berkeley.

Arakaki, A., Rodriguez Chamussy, L. y Vezza, E. (2020). Nowcasting Poverty in Argentina: A Methodological Note. Policy Research Working Paper. World Bank, Washington DC.

Bańbura, M., Giannone, D., Modugno, M., y Reichlin, L. (2013). Chapter 4: Now-casting and the real-time data flow. Graham, E., Timmerman, A. *Handbook of economic forecasting.* Elsevier.

Blanco, E., D'Amato, L., Dogliolo, F., y Garegnani, L. (2020). *Nowcasting Macroeconomic Aggregates in Argentina: Comparing the predictive ability of different models* (No. 4335). Asociación Argentina de Economía Política.

Blumenstock, J. E. (2016). Fighting poverty with data. *Science*, *353*(6301), 753-754.

Blumenstock, J. E. (2018, May). Estimating economic characteristics with phone data. In *AEA papers and proceedings* (Vol. 108, pp. 72-76).

Brum, M., y De Rosa, M. (2021). Too little but not too late: nowcasting poverty and cash transfers' incidence during COVID-19's crisis. *World Development*, *140*, 105227.

Cardinale Lagomarsino, B., Chagalj, C., y Romero, N. (2016). Predicción de la Pobreza en Argentina usando Random Forest.

D'Amato, L., Garegnani, L., y Blanco, E. (2015). *Nowcasting de PIB: evaluando las condiciones cíclicas de la economía argentina* (No. 2015/69). Working Paper.

Gasparini, L., Tornarolli, L., y Gluzmann, P. (2019). El desafío de la pobreza en Argentina. *Diagnósticos y Perspectivas*. CIPPEC-CEDLAS Working Paper.

Giannone, D., Reichlin, L., y Small, D. (2008). Nowcasting: The real-time informational content of macroeconomic data. *Journal of Monetary Economics*, *55*(4), 665-676.

Gonzalez-Rozada, M., y Menendez, A. (2002). Why have poverty and income inequality increased so much? Argentina 1991-2002. Centro de Investigaciones Financieras (CIF). Documentos de trabajo 01/2002.

Hyndman, R. J., y Khandakar, Y. (2008). Automatic time series forecasting: the forecast package for R. *Journal of statistical software*, *27*(3), 1-22.

Instituto Nacional de Estadísticas y Censos. (2003). La nueva Encuesta Permanente de Hogares de Argentina.

Instituto Nacional de Estadísticas y Censos. (2016). La medición de la pobreza y la indigencia en Argentina.

Lucchetti, L. (2018). *What can we (machine) learn about welfare dynamics from cross-sectional data?* The World Bank.

McBride, L., y Nichols, A. (2018). Retooling poverty targeting using out-of-sample validation and machine learning. *The World Bank Economic Review*, *32*(3), 531-550.

Observatorio de la Deuda Social Argentina, UCA (2020). La Medición de la Pobreza y el Problema de la Pobreza Estructural.

Parolin, Z., y Wimer, C. (2020). *Forecasting estimates of poverty during the COVID-19 crisis. Poverty and Social Policy Brief*. New York, NY: Center on Poverty and Social Policy at the Columbia School of Social Work, *4*(6).

Sampi Bravo, J. R. E., y Jooste, C. (2020). Nowcasting economic activity in times of COVID-19: An approximation from the Google Community Mobility Report. World Bank Policy Research Working Paper.

SEDLAC (2015). CEDLAS y Banco Mundial.

Silalahi, D. K. (2020). Forecasting of Poverty Data Using Seasonal ARIMA Modeling in West Java Province. *JTAM* (Jurnal Teori dan Aplikasi Matematika), 4(1), 76-86.

Sohnesen, T.P.y Stender, N. (2016). Is Random Forest a Superior Methodology for Predicting Poverty? An Empirical Assessment. Policy Research Working Paper;No. 7612. World Bank, Washington, DC.

Apéndice A. Pruebas de raíz unitaria

Tabla 2. Prueba de Raiz Unitaria

Variable	ADF	PP (zt)	KPSS	Estacionariedad
EMAE	-3,192*	-4,065***	0,316***	I(0)
Δ ln EMAE	-12,025***	-13,496***	0,0326	I(0)
Indice de Confianza del Consumidor	-3,485**	-3,583**	0,0592	I(0)
Salario Minimo	2,286	2,579	0,342***	I(1)
Δ ln Salario minimo	-15,271***	-15,611***	0,0563	I(0)
AUH	2,524	2,344	0,326***	I(1)
Δ ln AUH	-9,775	-13,338***	0,0722	I(0)
Evolucion del Empleo	-2,498	-2,167	0,329	I(1)
Δ ln Evolucion del Empleo	-3,197**	-6,549***	0,12*	I(0)
Plazo Fijo	-2,98	-3,024	0,285***	I(1)
Δ ln Plazo Fijo	-9,491***	-9,745***	0,355***	I(0)
Base Monetaria	4,176	4,138	1,97***	I(1)
Δ ln Base Monetaria	-12,862***	-12,790***	0,285***	I(0)
Depositos en Pesos	3,856	7,623	1,19***	I(1)
Δ ln Depositos en Pesos	-7,137***	-7,004***	0,102	I(0)
Depositos en USD	-1,37	-1,561	0,33***	I(1)
Δ ln Depositos en USD	-8,919***	-9,451***	0,141*	I(0)
Prestamos en Pesos	2,692	4,82	1,1***	I(1)
Δ ln Prestamos en Pesos	-5,866***	-5,698***	0,0655	I(0)
Prestamos en USD	-2,401	-0,823	0,207**	I(1)
Δ ln Prestamos en USD	-4,317***	-4,853***	0,142*	I(0)
Reservas Internacionales	-2,13	-1,917	0,603***	I(1)
Δ ln Reservas Internacionales	-10,586***	-10,602***	0,124*	I(0)
IPC Ecolatina	6,505	14,068	0,317***	I(1)
Δ ln IPC Ecolatina	-5,399***	-5,099***	0,0355	I(0)
Gasto en Planes	1,881	-4,653***	0,855***	I(1)
Δ ln Gasto en Planes	-17,060***	-50,507***	0,0106	I(0)
Tipo de Cambio Paralelo	1,847	1,59	0,702***	I(1)
Δ ln Tipo de Cambio Paralelo	-9,838***	-10,790***	0,0743	I(0)
Tipo de Cambio Oficial	5,679	5,96	1,84***	I(1)
Δ ln Tipo de Cambio Oficial	-9,445***	-9,441***	0,0441	I(0)
Remuneracion Privados	9,716	11,381	2,04***	I(1)
Δ ln Remuneracion Privados	-14,448***	-14,540***	0,0671	I(0)
Precio Soja Futuro	-2,291	-2,414	1,52***	I(1)
Δ ln Precio Soja Futuro	-15,067***	-15,050***	0,0674	I(0)
PIB	6,791	5,228	0,396***	I(1)
Δ ln PIB	-8,443***	-11,805***	0,045	I(0)
Consumo Nacional Privado	4,298	4,406	0,4***	I(1)
Δ ln Consumo Nacional Privado	-12,641***	-9,288***	0,0541	I(0)
Consumo Publico	11,941	5,996	0,406***	I(1)
Δ ln Consumo Publico	-16,668***	-30,841***	0,1	I(0)
Exportaciones FOB	0,339	1,092	0,34***	I(1)
Δ ln Exportaciones FOB	-9,49***	-10,155***	0,0807	I(0)
Tasa de Actividad	-5,117***	-5,09***	0,0503	I(0)
Tasa de Empleo	-4,368***	-4,244***	0,227***	I(0)
Δ ln Tasa de Empleo	-9,513***	-10,491***	0,0306	I(0)
Tasa de Desocupacion	-2,469	-2,143	0,393***	I(1)
Δ ln Tasa de Desocupación	-9,573***	-10,127***	0,035	I(0)

Apéndice B. Descripción de los datos

Tabla 3. Transformaciones sobre los Predictores Macroeconómicos

Dato	Frecuencia	Tipo	Transformación
Producto Bruto Interno (millones de pesos, a precios de 2004)	Trimestral	Flujo	No se le hizo transformación porque ya es trimestral
Estimador Mensual de Actividad Económica (Base 2004)	Mensual	Flujo	Promedio de los meses
Consumo Nacional Privado (millones de pesos, a precios corrientes)	Trimestral	Flujo	No se le hizo transformación porque ya es trimestral
Consumo Público (millones de pesos, a precios corrientes)	Trimestral	Flujo	No se le hizo transformación porque ya es trimestral
Exportaciones FOB (millones de pesos, a precios corrientes)	Trimestral	Flujo	No se le hizo transformación porque ya es trimestral
Indice Confianza del Consumidor (UTDT)	Mensual	Flujo	Promedio de los meses
Tasa de Actividad (Total Aglomerados)	Trimestral	Flujo	No se le hizo transformación porque ya es trimestral
Tasa de Empleo (Total Aglomerados)	Trimestral	Flujo	No se le hizo transformación porque ya es trimestral
Tasa de Desocupación (Total Aglomerados)	Trimestral	Flujo	No se le hizo transformación porque ya es trimestral
Salario Mínimo, Vital y Móvil (en pesos corrientes)	Mensual	Flujo	Promedio de los meses
Asignación Universal por Hijo (en pesos corrientes)	Mensual	Flujo	Promedio de los meses
Evolución del Empleo	Mensual	Flujo	Promedio de los meses
Tasa de Plazo Fijo 30 días (en % nominal anual)	Mensual	Flujo	Promedio de los meses
Base Monetaria (millones de pesos)	Mensual	Stock	Suma de los meses
Depósitos en Pesos (millones de pesos)	Mensual	Stock	Suma de los meses
Depósitos en Dólares (millones de USD)	Mensual	Stock	Suma de los meses
Préstamos al Sector Privado en Pesos (millones de pesos)	Mensual	Stock	Suma de los meses
Préstamos al Sector Privado en Dólares (millones de USD)	Mensual	Stock	Suma de los meses
Reservas Internacionales del BCRA (millones de USD)	Mensual	Stock	Suma de los meses
Gasto Público Nacional en Política de Ingresos (en millones de pesos corrientes)	Mensual	Flujo	Promedio de los meses
Tipo de Cambio Paralelo (USD)	Diario	Flujo	Promedio de los días
Tipo de Cambio Oficial (USD)	Diario	Flujo	Promedio de los días
Remuneración Promedio de los Trabajadores del Sector Privado (desestacionalizada)	Mensual	Flujo	Promedio de los meses
Precio Soja Futuro EEUU (USD)	Mensual	Flujo	Promedio de los meses