Prediciendo la Inflación Argentina.
Un enfoque probabilístico.

Tomás Marinozzi

Octubre 2022

Abstracto

Los forecast probabilísticos están ganando popularidad en la disciplina macroeconómica ya que los point forecasts no logran capturar el nivel de incertidumbre en variables fundamentales como inflación, crecimiento, tipo de cambio o desempleo. Este artículo explora el uso de modelos probabilísticos para predecir el índice de precios al consumidor de la Argentina. Se utilizaron scoring rules para evaluar diferentes modelos autorregresivos en relación a un benchmark. Los resultados muestran que los modelos univariados parsimoniosos tienen un rendimiento relativamente similar al de los modelos multivariados en torno a escenarios centrales, pero fallan en capturar los riesgos de cola, particularmente para los horizontes más largos.

Clasificación JEL: C13, C32, C53, E31

Palabras claves: Probability Forecast, Forecasting, Inflation Forecast, Continuous Ranked Probability Scores.

1. Introducción

Las predicciones sobre los eventos futuros (*forecasts*) desempeñan un rol cada vez más importante en la disciplina económica, dado que un mayor acceso a los datos, en conjunto con computadoras más veloces redujeron sustancialmente el costo de generar predicciones (Agrawal, Gans, y Goldfarb, 2018). En la macroeconomía, la implementación de técnicas de forecasting juega un rol importante en la toma de decisiones de política económica, así como en el proceso de anclaje de expectativas. Los bancos centrales utilizan predicciones como una forma de estimar el comportamiento futuro del sistema económico y evaluar la implementación de políticas económicas. Predicciones eficaces permiten a los hacedores de política, alinear expectativas e inducir una perspectiva *foward looking* a los mercados.

Hasta el día de hoy, la mayoría de los pronósticos "públicos" se presentan como *point forecast* (pronósticos basados en escenarios centrales). Los forecast probabilísticos, por otro lado, intentan cuantificar la incertidumbre que rodea la proyección de la variable objetivo. Este último método ha sido utilizado en otras disciplinas durante mucho tiempo, y junto con él se encuentran técnicas que permiten la evaluación de los forecast probabilísticos. Brier (1950), Winkler y Murphy (1968), Savage (1971) son algunos de los pioneros más reconocidos en la literatura sobre la construcción y evaluación de modelos probabilísticos. Aunque los forecast probabilísticos fueron originalmente desarrollados para predecir el clima, durante las últimas tres décadas, su popularidad y uso ha aumentado en disciplinas como las finanzas computacionales (Duffie y Pan, 1997) y la predicción macroeconómica (Garratt, Lee, Pesaran, y Shin, 2003). En finanzas en particular, el auge del *risk managment* ha acelerado la inclusión de estas técnicas como práctica estándar. En el ámbito de la macroeconomía, en términos generales, los forecast probabilísticos no son una práctica regular, pero su popularidad crece constantemente.

El objetivo de este artículo es explorar algunos modelos de predicción probabilística que podrían ayudar a cuantificar la incertidumbre alrededor de la inflación en Argentina. Un modelo de esta naturaleza podría ser útil para los responsables de la política económica ya que permite rankear escenarios posibles en base a su probabilidad de ocurrencia permitiendo planificar medidas contingentes de manera acorde. También podrían ser utilizados para mejorar la formulación de los contratos que son sensibles a las expectativas de inflación, como negociaciones salariales, tasas bancarias y cualquier tipo de decisiones de inversión.

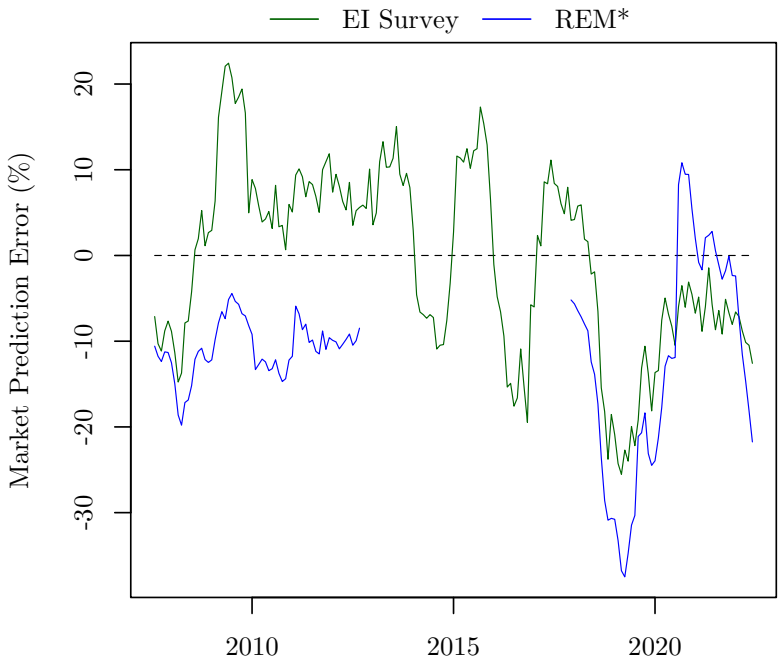
Este artículo está estructurado en seis secciones principales. La sección 2 analiza la necesidad de predicciones probabilísticas en un contexto como el de Argentina. La sección 3 proporciona una descripción de la estructura de los modelos que se utilizarán en el artículo, incluyendo un conjunto de diferentes modelos autorregresivos, el tratamiento de los shocks y el método para combinar forecast probabilísticos. La sección 4 discute las variables específicas y los tratamientos usados explícitamente para los diez modelos seleccionados. La sección 5 proporciona un entorno analítico sobre la estrategia de evaluación, describiendo algunas alternativas para evaluar forecast probabilísticos utilizando *scoring rules*, así como la evaluación de point forecast. La sección 6 ilustra los resultados del ejercicio de predicción, mientras que la sección 7 incluye algunos comentarios finales.

2. La necesidad de forecasts probabilísticos

El hecho de que se desconozca el "verdadero" modelo implica que cualquier pronosticador económico enfrenta incertidumbres y por ende tiene que aceptar cierto grado de imprecisión. En los últimos veinte años, la combinación de técnicas modernas de predicción y un poder computacional drásticamente superior, ha permitido a los pronosticadores simular cientos de escenarios para comprender mejor la naturaleza probabilística de la variable de interés. Sorprendentemente, a pesar del creciente poder computacional, los bancos centrales, el FMI, el Banco Mundial y otras instituciones de renombre mundial suelen publicar solamente los escenarios centrales (representados vía point forecasts). En algunos casos, algunas instituciones pueden incluso presentar un subconjunto simple de escenarios alternativos (favorable/optimista y desfavorable/pesimista), que carecen de aplicabilidad ya que las probabilidades de ocurrencia generalmente no son reveladas. A veces se presentan escenarios alternativos basados en cuantiles. Pero, en muchos casos, provienen de distribuciones ad-hoc cuya naturaleza predictiva no tiene necesariamente ningún fundamento empírico (por ejemplo, asumir que una distribución de shocks normal-invariante sea la distribución adecuada cuando, de hecho, raramente es el caso). Si bien los escenarios centrales (media o mediana, por ejemplo) pueden brindar una noción básica de direccionalidad y magnitud, brindan muy poca comprensión del grado de ocurrencia de tales eventos o los riesgos asociados a eventos extremos (riesgos de cola). Por lo tanto, uno puede pensar a los forecast probabilísticos como una manera de cuantificar la incertidumbre con el objetivo de tomar decisiones de manera más calificada.

Tomando como ejemplo la inflación en Argentina, caracterizada por altos niveles y particularmente

volátil durante los últimos quince años, se puede notar la fragilidad del point forecast como instrumento. Para ello, el siguiente gráfico compara el error de predicción de la inflación interanual derivado de la encuesta de expectativas de inflación de la Universidad Di Tella y el Relavamiento de Expectativas de Mercado (REM) publicado por el Banco Central de la República Argentina (BCRA).



(∗) REM no fue publicado entre Sep. 2012 - May. 2016

Figura 1: Errores de predicción - Inflación interanual. Fuente: UTDT & BCRA

En un caso como este, es evidente que el point forecast no brinda suficiente contexto para los riesgos inflacionarios. Independientemente del sesgo positivo o negativo en el error de predicción, el margen de error es sustancial. En la muestra se observa que el *mean average error* (MAE) es de alrededor de 10 puntos porcentuales. Es importante reconocer que esto probablemente se deba más a la volatilidad que exhibe la inflación que a "fallas" en la capacidad del mercado para predecir adecuadamente la inflación. Por lo tanto, se puede argumentar que si la inflación exhibe tales niveles de incertidumbre, entonces no se debe ignorar sino cuantificar. En algunos casos, los bancos centrales proporcionan los cuantiles derivados de las encuestas de inflación (como es el caso del REM). La figura (2) muestra la inflación real (12 meses rezagada) contra la media y los cuantiles (0.25 y 0.75) provenientes de la encuesta realizada por el Banco Central.

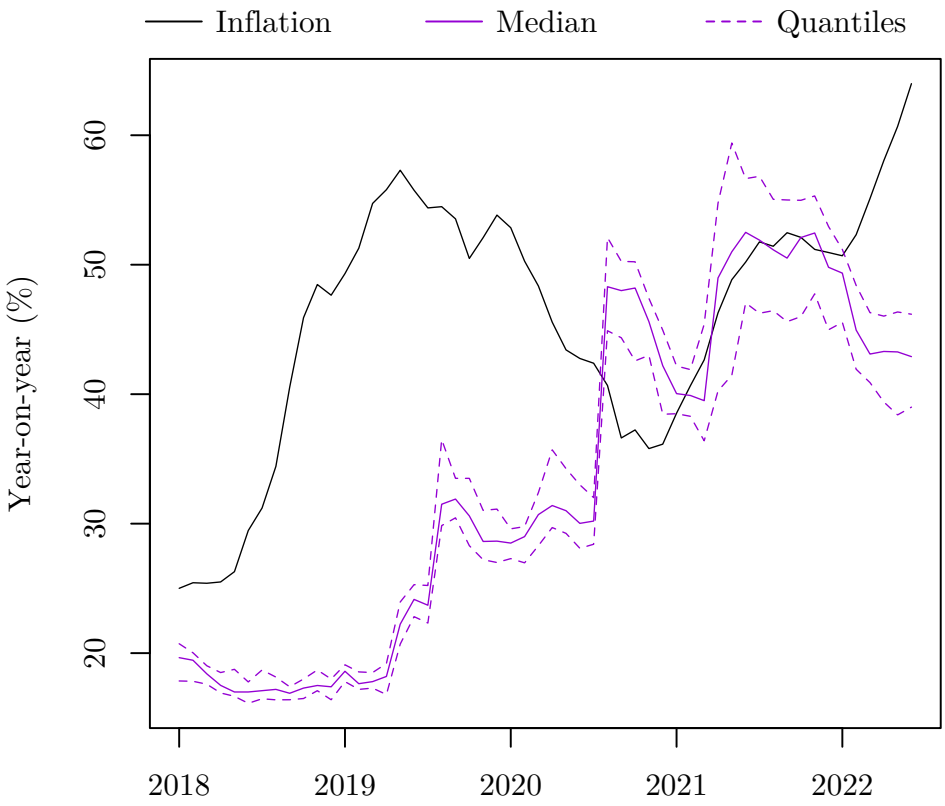


Figura 2: REM contra la inflación (12 meses rezagada). Fuente: BCRA

Notar cómo en algunos períodos los cuantiles son extremadamente estrechos respecto a la media (particularmente durante los primeros 24 meses del gráfico). Esto se debe a que una distribución de los distintos point forecast no necesariamente refleja la incertidumbre inflacionaria sino el grado de dispersión de las expectativas. Si bien es lógico suponer que un entorno más incierto generalmente se asocia con un mayor grado de dispersión en las expectativas, esto no captura adecuadamente los eventos de riesgo de cola. Para capturar adecuadamente los riesgos, se deben realizar ejercicios probabilísticos.

3. Estructura de los modelos

En línea con la sección anterior, se realizó un ejercicio de forecasting probabilístico con el objetivo de explorar modelos que capturen adecuadamente no solo los escenarios centrales sino también los riesgos inflacionarios para la Argentina. Naturalmente, cada modelo tiene sus propias ventajas y desventajas. Por ejemplo, se podría argumentar que los modelos univariados no pueden capturar completamente las interacciones que impulsan el sistema macroeconómico. Sin embargo, es posible que los modelos multivariados con un gran numero de variables y estructura posean un rendimiento pre-

dictivo inferior que aquellos modelos parsimoniosos, ya que pueden sufrir una sobre-especificación que conduzca al *overfitting*. Esta sección detalla las distintas estructuras autorregresivas selecionadas para lidiar con la media y algunos tratamientos para los shocks estocásticos.

3.1. Modelos de media condicional

3.1.1. Modelos univariados (benchmark)

Se probaron una serie de modelos univariados convencionales para predecir la inflación. En primer lugar, una caminata aleatoria (*random walk)*:

$$\pi_t = \pi_{t-1} + u_t \tag{1}$$

El random walk es quizás el punto de referencia más utilizado en la literatura macro y financiera, principalmente debido a su simplicidad y a su razonable capacidad de predicción. Sin embargo, también probamos otras especificaciones autorregresivas univariadas para tener una base alternativa.[1]

$$\pi_t = \rho_1 \pi_{t-1} + \cdots + \rho_p \pi_{t-p} + u_t \tag{2}$$

donde $u_t \sim N(0, \sigma)$[2].

3.1.2. Curva de Phillips

Se incluyó una versión de la "Hybrid New Keynesian Phillips Curve" (NKPC), propuesta originalmente por Galí y Gertler. La especificación elegida es una versión autorregresiva de la utilizada por D'Amato, Aguirre, Garegnani, Krysa, y Libonatti (2018) que contempla las características particulares de una economía abierta.

$$\pi_t = \phi_1 \pi_{t-1} + \phi_2 \mathrm{E}_{t-1}\left[\pi_t\right] + \delta x_{t-1} + \gamma \pi^*_{t-1} + \lambda \Delta e_{t-1} + u_t \tag{3}$$

En esta especificación, la devaluación del tipo de cambio, Δe_{t-1}, y la inflación internacional, π^*_{t-1}, tienen un efecto directo sobre la inflación doméstica (Svensson, 2000).

[1] Debido a la naturaleza no estacionaria de CPI, tratamos el modelo en variaciones porcentuales para modelos univariados e incluimos un término tendencial de ser necesario.

[2] Especificaciones alternativas para la varianza se discutirán más adelante en la sección.

3.1.3. Vector autorregresivo (VAR)

Pasando a los modelos multivariados, el modelo icónico de vectores autorregresivos (VAR) juega un papel importante en la literatura de predicción económica y financiera. Fueron introducidos por primera vez por Manz y Sims Jr (1980) como un método para analizar datos macroeconómicos y se hicieron populares debido a su simplicidad y su uso como una alternativa flexible a los modelos econométricos a gran escala.

Un conjunto de variables endógenas, $\boldsymbol{Y}_t$, son representadas como una función lineal de sus propios rezagos. Esto supone que las variables endógenas son tratadas simétricamente y que existe un efecto de retroalimentación entre ellas. No se descartó la posibilidad de regresores exógenos, $\boldsymbol{X}_t$, que pudieran afectar el comportamiento de la economía. Esto puede ser importante ya que nos enfrentamos a una pequeña economía abierta con una dinámica inflacionaria que depende de los precios internacionales de los commodities, los flujos de fondos y la actividad mundial.

El modelo VAR se puede expresar en su forma reducida como

$$\boldsymbol{Y}_t = \boldsymbol{\nu} + \boldsymbol{A}_1\boldsymbol{Y}_{t-1} + \ldots + \boldsymbol{A}_p Y_{t-p} + \boldsymbol{B}_1\boldsymbol{X}_{t-1} + \ldots + \boldsymbol{B}_q\boldsymbol{X}_{t-q} + \boldsymbol{u}_t \tag{4}$$

$\boldsymbol{Y}_t$ es un vector de dimensión $N \times 1$ de variables aleatorias endógenas; $\boldsymbol{X}_t$ es un vector de dimensión $L \times 1$ de variables exógenas; p y q son el número de rezagos para los vectores de variables endógenas y exógenas respectivamente; $\boldsymbol{\nu}$ es un vector fijo de constantes, de dimensión $N \times 1$; $\mathbf{A}_i$ son matrices de coeficientes de dimensión $N \times N$ para las variables endógenas; $\mathbf{B}_j$ son matrices de coeficientes de dimensión $N \times L$ para las variables exógenas; $\mathbf{u}_t \sim (\mathbf{0}, \boldsymbol{\Sigma_t})$ es un vector de dimensión $N \times 1$ de shocks exógenos serialmente no correlacionados ($\mathrm{E}\left[\mathbf{u}_t\mathbf{u}_s'\right] = \mathbf{0}\ \forall\ s \neq t$) con matriz de covarianza constante de dimensión $N \times N$ y media cero ($\mathrm{E}[\boldsymbol{u}_t] = 0\ \forall\ t$). Los supuestos anteriores implican una media condicional $\boldsymbol{\mu}_t$ y una covarianza constante $\boldsymbol{\Sigma_t} = \boldsymbol{\Sigma}$

Es relevante mencionar que se asume que las variables exógenas siguen un proceso separado de generación de datos. Este es un punto importante porque, para $h > 1$, tiene que haber una predicción paralela predeterminada para $\boldsymbol{X}_{t+h}$ afectando $\boldsymbol{Y}_{t+h}$. Esto podría ser un problema ya que la capacidad predictiva del regresor depende estrictamente de una predicción paralela. En pocas palabras, incluso si el "verdadero" valor de $\boldsymbol{X}$ tiene un contenido predictivo sustancial sobre $\boldsymbol{Y}$, si las proyecciones del regresor exógeno son de baja calidad, agregar un regresor exógeno podría empeorar la capacidad predictiva del modelo[3].

[3]En este escrito, las variables exógenas como el IPC de EE. UU. se modelaron por separado utilizando versiones

3.1.4. Vector of error correction (VEC))

Los modelos *vector of error correction* (VEC) son versiones restringidas de los modelos VAR diseñados con la intención de lidiar con series no estacionarias que están *cointegradas*. En esencia, los modelos VEC tienen especificaciones que restringen el comportamiento de largo plazo de las variables endógenas con el objetivo de forzar una convergencia a sus relaciones de largo plazo mientras se evidencian perturbaciones a corto plazo.

Para los modelos VEC, se realizó una test de Johansen para especificar el número de relaciones de cointegración y estimar estas relaciones. Dicho esto, seguir el enfoque de Johansen no es una condición necesaria ya que, en última instancia, los modelos se juzgan únicamente por su capacidad predictiva, pero utilizar un test econométrico puede ser buen punto de partida. Un VEC puede escribirse en forma matricial como,

$$\Delta \boldsymbol{Y}_t = \boldsymbol{\nu} + \boldsymbol{\Pi} \boldsymbol{Y}_{t-1} + \boldsymbol{\Gamma}_1 \Delta \boldsymbol{Y}_{t-1} + ... + \boldsymbol{\Gamma}_{(p-1)} \Delta \boldsymbol{Y}_{t-(p-1)} + \boldsymbol{u}_t \tag{5}$$

donde $\boldsymbol{\Pi} = -\left(\boldsymbol{I}_N - \boldsymbol{A}_1 - \cdots - \boldsymbol{A}_p\right)$ puede también puede escribirse como $\boldsymbol{\Pi} = \boldsymbol{\alpha}\boldsymbol{\beta}'$, donde $\boldsymbol{\beta}$ es la "matriz de cointegración" y $\boldsymbol{\alpha}$ representa la "matriz de carga"; $\boldsymbol{\nu}$ representa la tendencia determinista del proceso dinámico[4].

3.1.5. Imponiendo equilibrios de largo plazo

La estructura algebraica de los modelos VEC es bastante atractiva para aquellos que realizan proyecciones macroeconómicas, ya que es posible explicitar relaciones teóricas de largo plazo provenientes de la literatura. Esto es útil ya que existe mayor consenso sobre las relaciones de equilibrio a largo plazo, en lugar de la dinámica a corto plazo. Este enfoque ha sido utilizado anteriormente por otros autores (Garratt, Lee, Hashem Pesaran, y Shin, 2003; Schneider, Chen, y Frohn, 2008) y se los conoce como modelos *Long-Run*.

El documento aprovecha la estructura del modelo VEC para probar algunos modelos que incluyen teorías muy discutidas en la literatura que podrían (potencialmente) mejorar el rendimiento de los modelos de predicción. Probamos dos modelos con teorías de largo plazo. La primera incluye dos

testeadas de modelos univariados.

[4]Leer Lütkepohl (2005) sobre las diferentes formas de modelar la tendencia determinística, así como la versión con regresores exógenos.

relaciones de tipo de cambio, *paridad del poder adquisitivo* (PPP) y la *paridad de la tasa de interés descubierta* (UIP),

$$\begin{aligned} PPI: &\quad P_t = E_t - P_t^* \\ UIP: &\quad \Delta E_t = i_t - i_t^* \end{aligned} \tag{6}$$

Donde P_t es el logaritmo del precio doméstico, P_t^* es el logaritmo del precio internacional, E_t es el logaritmo del tipo de cambio, i_t es la tasa de interés doméstica y i_t^* es la tasa de interés internacional.

El segundo modelo incluye *neutralidad del dinero* (MN) y *equilibrio en el salario real* (RWE).

$$\begin{aligned} MN: &\quad M_t - P_t = k \\ RWE: &\quad W_t - P_t = \delta \end{aligned} \tag{7}$$

donde M_t y W_t representan la oferta monetaria y los salarios nominales. k y δ representan constantes que guían las relaciones a largo plazo[5].

3.2. Volatilidad condicional e innovaciones no paramétricas

Hasta ahora, se ha asumido un proceso de ruido blanco paramétrico para los residuos, es decir, media cero con varianza constante. El documento también explora distribuciones no paramétricas basadas en técnicas de remuestreo, así como modelos de volatilidad condicional. Se podría argumentar que, en algunos casos, es más probable que los entornos con alta volatilidad permanezcan constantes hasta que se presente un cambio en las condiciones económicas o el régimen político, lo que podría causar heteroscedasticidad o también *clusters* de volatilidad.

La clave detrás de este tipo de modelos es que σ_t^2 está condicionado por la información pasada $\mathcal{F}_{t-1}$. Asumiendo que z_t es un ruido blanco con media cero y varianza unitaria constante, y la varianza condicional σ_t^2 se expresa como

$$u_t = \sigma_t z_t$$

[5]Se podría discutir la validez de k y δ como constantes. En este ejercicio en particular, dada la ventana de datos escogida, los modelos se probaron con términos constantes pero la metodología utilizada permite un proceso dinámico para esas variables.

donde z_t es una secuencia de variables aleatorias independientes e idénticamente distribuidas con media cero y varianza unitaria.

3.2.1. Innovaciones GARCH

Los modelos autorregresivos de heteroscedasticidad condicional (ARCH) describen la varianza actual como una función del cuadrado de los términos de error de los períodos anteriores. La versión inicial de esta familia de modelos fue desarrollada por Engle (1982), evolucionando rápidamente a una versión generalizada (GARCH) introducida por primera vez por Bollerslev (1986). La versión generalizada incluye el proceso ARCH con versiones de rezagos adicionales en la varianza. Siguiendo la especificación GARCH convencional, se asume que la heteroscedasticidad condicional está dada por:

$$\sigma_t^2 = \gamma + \sum_{i=1}^{m} \alpha_i u_{t-i}^2 + \sum_{j=1}^{s} \beta_j \sigma_{t-j}^2 \tag{8}$$

bajo el supuesto de que m and s son enteros no negativos, donde $\gamma > 0, \alpha_i \geq 0, \beta_j \geq 0$ para todo $i > 0$, y $j > 0$ y $\sum_{i=1}^{m} \alpha_i + \sum_{j=1}^{s} \beta_j \leq 1$. El componente ARCH del modelo se describe como $\sum_{j=1}^{s} \beta_j \sigma_{t-j}^2$.

Un problema al modelar la varianza condicional, es que la cantidad de parámetros aumenta y, en última instancia, podría llevar a un overfitting, particularmente en muestras cortas como en este caso. Para evitar un numero notorio de parámetros, utilizamos una especificación GARCH(1,1). Este supuesto está parcialmente respaldada por evidencia empírica. Hansen y Lunde (2005) compararon 330 modelos de volatilidad condicional, en su caso usando datos diarios, y llegaron a la conclusión de que no hubo una mejora significativa al usar un modelo de predicción diferente a GARCH(1,1). Aunque el estudio original se centró en la volatilidad de las acciones, esta especificación en particular es una práctica estándar para los pronosticadores.

Es posible derivar una extensión multivariante del modelo GARCH (MGARCH), para permitir que la matriz de covarianza de las variables dependientes siga una estructura dinámica condicionada por la información pasada. Sin embargo, para este análisis específico, este tipo de modelos no se consideran ya que el número de parámetros crece exponencialmente con el número de variables. Se intentó una variación de este tipo de modelos multivariados pero debido a la baja frecuencia de los datos (mensual en lugar de diaria) y la poca historia de la muestra, el uso de modelos MGARCH se descartó rápidamente debido al overfitting y un rendimiento extremadamente pobre, particularmente en los primeros períodos del out-of-sample. Estos tipos de modelos GARCH multivariados no se incluirán en el análi-

sis, entonces, para preservar la estructura covariante en modelos multivariados, además de un proceso GARCH, se utiliza bootstrapping para garantizar que los errores GARCH simulados conserven una distribución conjunta.

3.2.2. Innovaciones Bootstrap

Bootstrapping es otra técnica utilizada en este artículo para introducir innovaciones. En este caso, las innovaciones son derivadas mediante distribuciones no paramétricas. En este caso, los shocks salen de una distribución generada a través de un remuestreo de los residuos del modelo. La metodología utilizada es similar a la propuesta originalmente por (Efron, 1992) Efron.

Sea X una tupla de residuos multivariados tales que $X = \{u^1, \cdots, u^s\}$, donde s representa el número de variables en el modelo. Para unos residuos $X = \{X_1, \cdots, X_n\}$, donde n es el número de tuplas disponibles, el ejercicio consiste en un remuestreo aleatorio de X creando una matriz de innovaciones de tamaño $s \times r \times h$ donde r representa el número de simulaciones y h el número de horizontes.

3.3. Mixture de modelos

En general, es posible reducir la incertidumbre del modelo utilizando una combinación de modelos. En este caso, una combinación de distribuciones probabilísticas, o lo que es lo mismo un *mixture* de distribuciones. Para representar esto matemáticamente, se asume que la variable objetivo y es generada por una variable latente z no observada. Formalmente, $p(z)$ es una distribución multinomial, mientras que $p(y|z)$ puede tomar una variedad de formas paramétricas. Podemos calcular la función de densidad de probabilidad sobre y al marginalizar z de la siguiente manera

$$p(y) = \sum_{i=1}^{K} P[Z = z_i]\, p(y|z = z_i)$$

Es importante distinguir la diferencia entre un mixture de distribuciones y un promedio ponderado de las distribuciones. En la práctica, promediar dos distribuciones de igual tamaño corresponde a un promedio ponderado componente por componente[6]. Un mixture, por otro lado, extrae una muestras z_i de las distintas distribuciones con la frecuencia $p(z)$. ¿Por qué es importante hacer esta distinción? Supongamos que dos modelos probabilísticos predicen las siguientes distribuciones (Gaussianas) para la misma variable aleatoria y (vea la figura (3) a continuación). Obsérvese que al promediar ambas distribuciones, la distribución resultante sigue siendo Gaussiana, mientras que el mixture genera una

[6] Asumiendo distribuciones de misma dimensión.

distribución no Gaussiana, contemplando aspectos de ambas distribuciones.

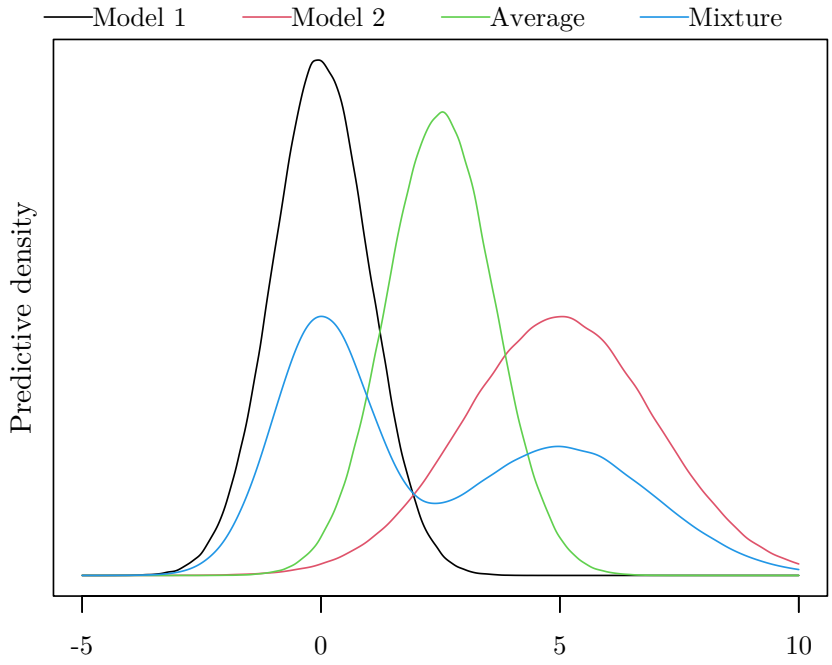


Figura 3: Distribuciones arbitrarias.

Conceptualmente, suponiendo que ambas distribuciones provinieran de pronosticadores diferentes (F_1 y F_2), se podría argumentar que los pronosticadores no están de acuerdo sobre el resultado más probable. Mientras que F_1 dice que Y tomará un valor más cercano a cero, F_2 dice que estará más cerca de 5. Si se promediaran ambas distribuciones, el resultado sería la distribución intermedia (densidad verde) con la mayor probabilidad resultados alrededor de 2.5. Sin embargo, incluso cuando ambos pronosticadores no están de acuerdo con el valor esperado (la media), ambos están de acuerdo en que las probabilidades de que la variable y resulte ser alrededor de 2.5 son bajas. Es por eso que para los forecast probabilísticos, promediar las distribuciones puede no ser necesariamente el mejor enfoque, ya que puede asignar una alta probabilidad a escenarios realísticamente poco probables. Entonces, este artículo se enfocará en mixture de modelos como método para combinar modelos.

4. Modelos seleccionados y variables incluidas

Como se mencionó en la sección 1, se probaron treinta modelos diferentes. Sin embargo, con fines ilustrativos, solo se mostrará una selección de diez modelos (además del benchmark). Esta sección describe los modelos escogidos y las variables incluidas en aquellos modelos. El conjunto de datos empieza en febrero de 2004 y finaliza en diciembre de 2019. Las variables incluidas son:

- IPC: Índice de Precios al Consumidor para la Argentina[7].
- Expectativas: Expectativa de inflación a doce meses (media). Fuente: Universidad de Di Tella.
- EMAE: Estimador Mensual de Actividad Económica. Fuente: INDEC.
- Tasa de interés: Tasas de interés por depósitos a plazo fijo de 30 a 35 días de plazo. Fuente: Banco Central de la República Argentina (BCRA).
- Salarios: Salario medio de los trabajadores registrados del sector privado. Fuente: Ministerio de Trabajo.
- Dinero: Una aproximación sumando la base monetaria y los pasivos a corto plazo del banco central. Fuente: Banco Central de la República Argentina (BCRA).
- ARS/USD: Tipo de cambio nominal bilateral. Fuente: Banco Central de la República Argentina (BCRA).
- IPC de EE. UU.: Índice de precios al consumidor de EE. UU. Fuente: Oficina de Análisis Económico de EE. UU. (BEA).
- Tasa de interés de EE. UU.: Tasa de mercado de un letra del Tesoro a 3 meses. Fuente: Junta de la Reserva Federal.

El cuadro 1 aclara cuáles fueron los modelos seleccionados, las variables utilizadas y el tratamiento de varianza específico para generar los shocks aleatorios detrás de las simulaciones. Por razones obvias, todos los modelos usan el IPC y, por lo tanto, la variable se excluyó del cuadro[8]. El modelo Long-Run del cuadro incluye neutralidad del dinero y salarios reales constantes, el Long-Run alternativo, que incluye PPP y UIP, no se incluyó debido a su menor desempeño.

[7] El índice de precios se construyó combinando el índice del Instituto Nacional de Estadística y Censos de Argentina (INDEC) y el Índice de Precios de la Ciudad de Buenos Aires y de San Luis. La metodología es idéntica a la utilizada por la Universidad del CEMA. (ver https://ucema.edu.ar/cea_vce/serie).

[8] Vale aclarar que la tasa de interés de EE. UU. también se excluyó del cuadro porque ninguno de los modelos seleccionados la utiliza.

	Tratamiento de Varianza	Expectativas	EMAE	ARS/USD	Salarios	Oferta monetaria	Tasa de interés	U.S. CPI
(0) RW	Garch (1,1)							
(1) AR(1)	Paramétrico							
(2) AR(2)	Garch (1,1)							
(3) AR(4)	Paramétrico							
(4) VAR(2)	Garch (1,1)		X		X		X	
(5) VAR(2)	Paramétrico		X		X		X	
(6) VEC(4)	Garch (1,1)			X	X	X		
(7) VEC(4)	Bootstrap			X	X	X		
(8) PC	Bootstrap	X	X	X				X
(9) Long-Run	Bootstrap	X	X		X	X		
(10) Mixture	-	X	X	X	X	X	X	X

Cuadro 1: Modelos y variables incluidas

El número de la izquierda representa el número de identificación de los modelos. A partir de ahora, referirse a los modelos por el nombre o por su número de identificación será indiferente. Se probaron un par de combinaciones, pero el mixture seleccionado fue la combinación de los modelos multivariados 6, 8, 9. Los modelos restantes y las variables utilizadas se muestran en la Tabla 4 dentro del apéndice.

5. Estrategia de evaluación

Con el objetivo de comparar los modelos, se realizó un *out-of-sample testing* (evaluación por fuera de la muestra). Los parámetros fueron estimados recursivamente sobre la etapa out-of-sample usando todas las observaciones disponibles hasta el momento de la predicción (Rossi, 2014). Con el objetivo de tener una evaluación más "realista" o "justa" vale la pena hacer ciertas aclaraciones de cómo se llevó a cabo el proceso recursivo frente el tratamiento de los datos faltantes y las revisiones de las series.

Respecto al primero punto, asumiendo que el pronosticador corre el modelo a final del mes, a la hora de hacer el enfoque recursivo se toma en cuenta cuáles datos en promedio tiene algún tipo de rezago debido a su calendario de publicación (ejemplos: EMAE, salarios, expectativas, etc). Para aquellas variables se les imputa el valor faltante utilizando un "kalman smoother" basado en un modelo estructural y estimado por máxima verosimilitud. Por otro, el IPC es una variable rezagada en si misma

con respecto a otras variables que tienen una frecuencia más alta (como por ejemplo el tipo de cambio o las tasas de interés). Para lidiar con aquel inconveniente el modelo corta la data recursivamente en la fecha del último valor del IPC realizado y calibra el shock tal que el resultado de las variables adelantadas replique al valor realizado en el primer horizonte. Esa fue la manera escogida para lidiar con el descalce de datos, no obstante, se reconoce que no se he hecho ningún out-of-sample testing sobre el método de imputación para los valores faltantes. Sobre este tema, Zanfei, Menapace, Brentan, y Righetti (2022) reconocen que diferentes métodos de imputación generan diferencias sustanciales en la calidad de las predicciones.

Otro aspecto importante es la revisión de las series estadísticas o el cambio de metodología. Es posible que los modelos sean sensibles a las revisiones de las series, entonces, para ser justo debería usarse la serie que originalmente estaba disponible en aquel momento, con la metodología en aquel momento. En este trabajo puntual un caso muy claro es el EMAE donde la serie tu cambios en su metodología y constantemente tiene revisiones en los últimos datos. Debido a la dificultad de encontrar todas las versiones anteriores de las series del EMAE, IPC e índice de salarios, este punto se ignora para este trabajo, pero merece ser aclarado ya que es pertinente debido a que los resultados están condicionados al conjunto de datos escogidos (Check, Nolan, y Schipper, 2018).

Finalmente, se aplicaron diferentes medidas de precisión para evaluar la capacidad predictiva de los modelos. Para los forecasts probabilísticos, se utilizó el *Continuous Ranked Probability Score* (CRPS) y el *Quantile Score* (QS) mientras que para los point forecasts (representado por la mediana del forecast probabilístico) se utilizó *Root Mean Square Error* (RMSE) y el *Mean Percentage Error* (MPE) para verificar el sesgo del modelo. Se realizó el test de inferencia de *Diebold-Mariano* (DM test) para evaluar si existía un modelo superador para la muestra. Finalmente, se utilizó el enfoque la transformada integral de probabilidad (PIT) para evaluar si el mixture de modelos estaba correctamente especificado.

5.1. Evaluación de point forecast

Para hacer una evaluación en out-of-sample, es necesario introducir algún tipo de métrica que mida el desempeño de los modelos y permita la comparación entre ellos. Esto implica que, para un horizonte determinado, h, existe una función de pérdida que mapea la desviación de la predicción contra los valores realizados a lo largo de la ventana del out-of-sample. Formalmente, para un modelo específico y una función de pérdida L, el desempeño promedio Π se define como

$$\Pi_h = \frac{1}{T}\sum_{t=1}^{T} L(\hat{y}_{t+h}, y_{t+h}) \tag{9}$$

donde $\hat{y}_{t+h}$ es la predicción producida h períodos atrás, mientras que y_{t+h} es el valor realizado. La función de pérdida más convencional para evaluar point forecasts es *Root Mean Square Error* (RMSE),

$$RMSE_h = \sqrt{\frac{1}{T}\sum_{t=1}^{T} (\hat{y}_{t+h} - y_{t+h})^2} \tag{10}$$

Mean percentage error (MPE) también se utilizará para visualizar si el modelo exhibe un sesgo en el horizonte h,

$$MPE_h = \frac{1}{T}\sum_{t=1}^{T} \frac{\hat{y}_{t+h} - y_{t+h}}{\hat{y}_{t+h}} \tag{11}$$

debido a que en la fórmula no hay términos cuadráticos (o en valores absolutos), ante la ausencia de sesgo, los errores de predicción positivos y negativos deberían en promedio compensarse[9].

5.2. Evaluación de forecasts probabilísticos

Análogo al concepto de función de pérdida, las *scoring rules* son técnicas para evaluar forecasts probabilísticos.

Definition 1. ***(Scoring rule)*** *Dada una función de distribución acumulativa (CDF) predicha por un pronosticador, $F \in \mathcal{F}$, para una variable aleatoria Y, la scoring rule S es un mapa tal que $S : \mathcal{F} \times \mathbb{R} \to \mathbb{R}$. Específicamente, la scoring rule asigna una puntuación numérica $S(F, y) \in \mathbb{R}$ a F después de evaluar su desempeño en relación con la observación real y.*

Al igual que con los point forecasts, las scoring rules tratan de minimizar el error esperado. Suponga que el agente cree que la verdadera distribución es G, entonces la puntuación esperada debería ser

$$\min_F \mathrm{E}_G S(F, y) = \min_F \sum_y q(y) S(F, y)$$

donde q representa una probabilidad. En este contexto, es importante reconocer scoring rules "justas" que premian a los pronosticadores que buscan la distribución real.

[9]Por último, notar que al usar CPI como variable a evaluar, no hay que preocuparse del denominador ya que la tendencia positiva implica que $\hat{y}_{t+h} > 0$.

Definition 2. ***(Scoring rules apropiadas)*** *Una scoring rule S es apropiada (con respecto a la clase $\mathcal{F}$) si la pérdida esperada es minimizada en la verdadera CDF, es decir, si $Y \sim G$. Entonces*

$$\mathbb{E}_G S(G, Y) \leq \mathbb{E}_G S(F, Y), \quad \forall F \in \mathcal{F}$$

Una scoring rule es **estrictamente apropiada** si su valor esperado es *minimizado únicamente* por la verdadera distribución de probabilidad. Deben evitarse las reglas que no sean apropiadas, ya que podrían alentar al pronosticador a presentar predicciones que él reconoce que son incorrectas. Se puede encontrar una revisión detallada de este tema en Gneiting y Raftery, (2007) y Bröcker y Smith, (2007).

5.2.1. Probability Score (BS)

El Probability (o Brier) score (BS), fue introducido por primera vez por Brier (1950), es un tipo de scoring rule que evalúa la precisión de la predicción en función de la distancia euclidiana entre la probabilidad real de una observación binaria (alrededor de un umbral) y la probabilidad predicha. En términos simples, el Brier score, muestran la capacidad del modelo para capturar la verdadera probabilidad de ocurrencia de un evento. Formalmente, el Brier score de un evento discreto $Y \in A$ con $p = P_F[Y \in A]$ se caracteriza por,

$$BS^A = (p - \mathbb{I}[y \in A])^2$$

donde $\mathbb{I}$ es una distribución binaria $[0, 1]$ que asigna probabilidad 0 a los eventos que no ocurrieron y 1 a los que sí. En el contexto de los modelos probabilísticos, se intenta encontrar modelos que logren capturar mejor la probabilidad de $y \leq z$ donde z es un umbral arbitrario. El mapeo de un evento discreto a una distribución es relativamente sencillo ya que cualquier predicción de densidad f induce un forecast probabilístico para el evento binario $Y \leq z$ a través de la CDF, es decir

$$F(z) = \int_{-\infty}^{z} f(y)dy$$

Por lo tanto, el Brier score se puede reescribir como

$$BS^z_{t,h}(F_{t+h}(z), y_{t+h}) = (p_{t+h} - \mathbb{I}[y_{t+h} \leq z])^2 \tag{12}$$

donde $p_{t+h} = P[Y_{t+h} \leq z] = F_{t+h}(z)$ para todo $t = 0, 1, 2, \ldots, T$. Para el caso predicciones unidimensionales, el Brier score es la transformación probabilística del error cuadrático que se usa para la evaluación de los point forecasts.

5.2.2. Quantile score (QS)

Si F es una función de distribución acumulativa monótonamente creciente, entonces es posible definir una función inversa única F^{-1}, a menudo denominada función *cuantil*. Las funciones de cuantiles permiten a los pronosticadores evaluar el desempeño de la distribución predictiva en los cuantiles (esto es particularmente relevante cuando se evalúa la capacidad de un modelo para predecir riesgos de cola). Para este propósito, la scoring rule más convencional es el *Quantile score* (QS)[10] (Koenker y Bassett Jr, 1978). Formalmente, el Quantile score se define como

$$\mathrm{QS}^{\alpha}_{t,h}(F^{-1}_{t+h}(\alpha), y_{t+h}) = 2(\mathbb{I}\{y_{t+h} < q\} - \alpha)(q - y_{t+h}) \tag{13}$$

donde $q = F^{-1}_{t+h}(\alpha)$ para un cuantil $\alpha \in (0, 1)$.

5.2.3. Continuous ranked probability score (CRPS)

Las scoring rules analizadas hasta ahora evalúan una porción específica de la distribución, ya sea una región de probabilidad o un cuantil específico de la distribución. El *Continuous ranked probability score* (CRPS)[11] permite a los pronosticadores evaluar el desempeño predictivo de la distribución en su conjunto. Formalmente,

$$\mathrm{CRPS}_{t,h}(F_{t+h}, y_{t+h}) = \int_{-\infty}^{\infty} (F_{t+h}(x) - \mathbb{I}\{y_{t+h} \leq x\})^2 \, \mathrm{d}x \tag{14}$$

en el contexto de CRPS, $\mathbb{I}$ es una función escalonada de Heaviside que toma el valor 0 para cualquier valor por debajo del valor real y 1 para cualquier valor igual o superior al valor real (Matheson y Winkler, 1976).

También se puede dividir la integral original en dos, en el umbral crítico $y_{t+h} = x$ para simplificar la función escalón de Heaviside,

[10] También conocida como puntuación pinball, o puntuación lineal asimétrica por partes

[11] Esto a veces se denomina *distancia de error euclidiana estocástica* (Diebold y Shin, 2017).

$$\mathrm{CRPS}_{t,h}(F_{t+h}, y_{t+h}) = \int_{-\infty}^{y_{t+h}} F_{t+h}(x)^2 dx + \int_{y_{t+h}}^{\infty} (F_{t+h}(x) - 1)^2 dx$$

En la práctica, debido a que F_{t+h} es una distribución empírica, solo hay un número finito de puntos para evaluar, lo que significa que las integrales se pueden convertir en sumas discretas y finitas que son computacionalmente factibles.

Finalmente, es importante resaltar que existe un fuerte vínculo entre las tres scoring rules discutidas hasta ahora. De hecho, los primeros dos son equivalentes al CRPS cuando se agregan a lo largo de la distribución. Formalmente,

$$\mathrm{CRPS}_{t,h}(F_{t+h}, y_{t+h}) = \int_{-\infty}^{\infty} BS_{t,h}^{z}(F_{t+h}(z), y_{t+h})\, \mathrm{d}z = \int_{0}^{1} \mathrm{QS}_{t,h}^{\alpha}(F_{t+h}^{-1}(\alpha), y_{t+h})\, \mathrm{d}\alpha \tag{15}$$

Una vez obtenido el score de cada periodo del out-of-sample, se promedió el rendimiento de la scoring rule S generando una puntuación media de Π,

$$\Pi_h = \frac{1}{T} \sum_{t=1}^{T} S_{t,h}$$

5.3. Testeo del rendimiento predictivo

Para una función de pérdida dada, se puede testear formalmente si dos modelos rivales (por ejemplo, i y j) tienen el mismo desempeño predictivo usando el test de *Diebold-Mariano* (DM test). La prueba formal puede basarse en el estadístico,

$$t_h = \sqrt{T}\, \frac{\Pi_h^i - \Pi_h^j}{\hat{\sigma}_h^2} \tag{16}$$

donde

$$\hat{\sigma}_h^2 = \frac{1}{T} \sum_{i=1}^{T} \left(S_{t,h}^i - S_{t,h}^j \right)^2 \tag{17}$$

es una estimación de la varianza del diferencial de la scoring rule. El DM test no requiere de un comportamiento específico para los scores individuales, sin embargo, requiere que el diferencial de scores sea estacionario.

5.4. Calibración utilizando los PITs

Los métodos de evaluación descriptos hasta ahora solo son útiles para la comparación relativa contra algún benchmark (u otros modelos) ya que no existe una medida estándar de un valor CRPS "apropiado". Para proporcionar una medida de evaluación "absoluta" en lugar de "relativa", el pronosticador suele realizar un análisis de calibración basado en el uso de la *transformada integral de probabilidad* (PIT) (Diebold, Gunther, y Tay, 1997).

Una transformada integral de probabilidad (PIT) es la probabilidad acumulada evaluada en el valor realizado de la variable objetivo. Mide la probabilidad de observar un valor menor que el valor realizado, donde la probabilidad se mide utilizando la CDF. De acuerdo con (Diebold y cols., 1997), un forecast probabilístico está correctamente especificado si 1) los PITs se distribuyen uniformemente en el intervalo $(0, 1)$, 2) para predicciones one-step-ahead[12] los PITs son independientes (lo que significa que no hay autocorrelación).

6. Resultados

Como se mencionó en las secciones anteriores, se testearon treinta modelos diferentes utilizando una estimación recursiva out-of-sample para doce horizontes. La evaluación comienza en enero de 2012 hasta diciembre de 2019, dividiendo la data en dos partes, aproximadamente la mitad para la estimación in-sample y la otra mitad para la estimación out-of-sample. Con fines ilustrativos, se seleccionaron diez modelos (más el benchmark) para incluirlos en los gráficos y cuadro[13].

Para una variable objetivo como la inflación, no es obvio qué transformación del índice es más apropiada para la evaluación. Por ejemplo,Koop y Korobilis (2011) utiliza inflación mensual, Croushore y Van Norden (2018) usa inflación trimestral anualizada, Stock y Watson (2008) utiliza inflación interanual y Riofrío, Chang, Revelo-Fuelagán, y Peluffo-Ordóñez (2020), Zahara y cols. (2020), hacen la evaluación directamente sobre el IPC. Este documento evalúa el desempeño del índice de precios en sí mismo en lugar de una transformación específica de los datos[14]. Dicho esto, es muy común discutir la evolución del índice a doce meses, entonces, se prestará especial atención al horizonte doce ($h = 12$) cuando se muestren algunos gráficos de horizonte fijo.

[12] En la práctica, el pronosticador tiende a probar la calibración en las predicciones one-step-ahead. Si bien hay literatura sobre predicciones multi-step-ahead, producen errores de predicción auto correlacionados y por ende PITs autocorrelacionados, lo cual complica el análisis (Knüppel, 2015). La calibración de la predicción de multi-step-ahead va más allá del alcance de este artículo.

[13] Los resultados de CRPS para el resto de los modelos se pueden encontrar en el cuadros 5 dentro del apéndice.

[14] Debido a que el IPC no es estacionario, los datos se transformaron en diferencias porcentuales o diferencias logarítmicas durante la estimación y el proceso predictivo, pero luego la proyección se revirtió al IPC para comparar.

6.1. Testeo en out-of-sample

En general, las métricas, incluyendo Continuous Ranked Probability Scores (CRPS), Quantile Scores (QS), Root Mean Square Errors (RMSE), se muestran en términos relativos respecto al benchmark. Los lectores deben tener en cuenta que se desea un CRPS más bajo, por lo tanto, un rendimiento relativo más bajo con respecto al benchmark implica ganancias predictivas contra el benchmark[15]. También se utilizó el Mean Percentaje Error (MPE) para verificar el sesgo en los point forecasts[16].

La figura (4) muestra el rendimiento CRPS relativo de los modelos. Obsérvese que hay un subconjunto de modelos que, para la ventana de datos en out-of-sample, superaron al benchmark (es decir un CRPS relativo más bajo) en todos los horizontes, en contraste con otros modelos que solo superaron al benchmark en algunos horizontes. Por ejemplo, el modelo de Long-Run seleccionado, que tiene dos relaciones de largo plazo y obtuvo un rendimiento inferior al benchmark en horizontes más cortos (horizontes 1 a 5) pero superior en horizontes más lejanos.

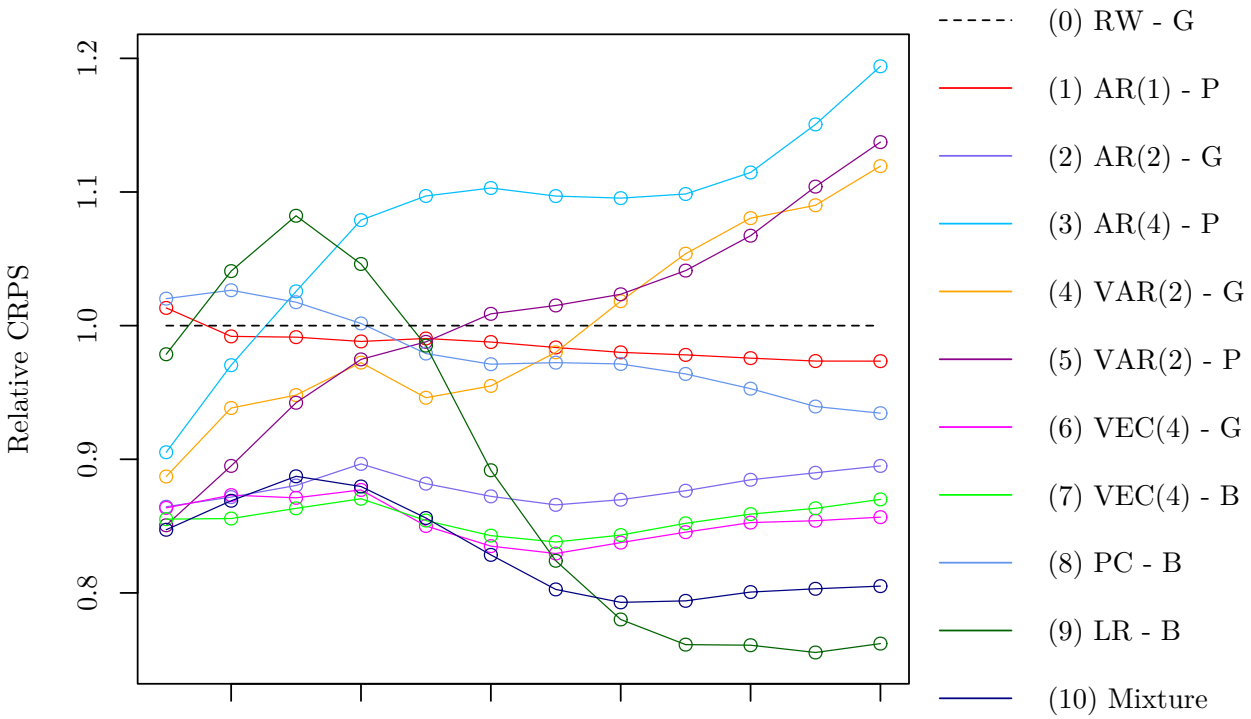


Figura 4: CRPS relativo por horizonte

[15]Es decir, se desea un CRPS / QS / RMSE relativo más bajo.

[16]En este caso el rendimiento no se compara contra el benchmark, ya que el benchmark podría haber tenido sesgo, contaminado el análisis, entonces una comparación relativa no es apropiada.

Por otro lado, los modelos VAR seleccionados tienden a tener un desempeño superior en horizontes más cortos pero no logran capturar la dinámica a largo plazo. En general, los modelos de rendimiento superior tienen ganancias de entre un 5 % y un 15 % en relación al Random Walk, mientras que el modelo de Long-Run tiene el mejor rendimiento de todos los modelos en los horizontes 9 a 12 (entre 20-25 % por encima del benchmark).

La figura (5) muestra el Quantile score para el horizonte 12. Resulta interesante que los modelos exhiben, en general, un desempeño similar en la mediana pero hay grandes diferencias en las colas.

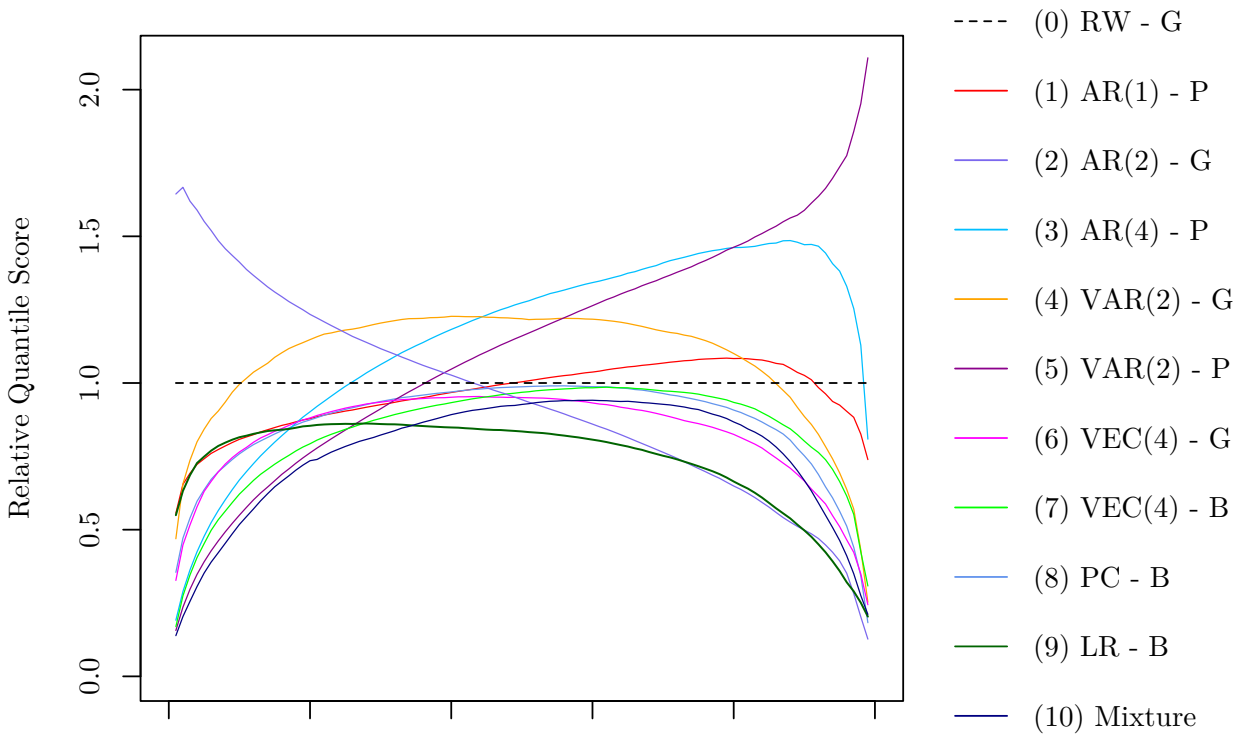


Figura 5: QS por cuantiles para el horizonte $h = 12$

Por ejemplo, los modelos 3 y 4 tienen un mejor desempeño en la cola izquierda, mientras que otros tienen un mejor desempeño solo en la cola derecha (riesgos de alta inflación), como el modelo 2. Algunos de los modelos tienen un desempeño similar en la mediana, pero superior en ambas colas, esto es importante ya que es un indicio que hay modelos que son particularmente mejores prediciendo riesgos inflacionarios (o mas genérico, eventos extremos).

La figura (6) muestra los diferenciales acumulados de CRPS. Esta métrica ayuda a comprender la evolución del rendimiento del modelo (en relación con el benchmark) a lo largo de la muestra. Cabe señalar que la métrica no se puede expresar como un porcentaje debido a que en los primeros

períodos los puntajes CRPS acumulados son aproximadamente cero, lo que provoca inestabilidad en el cociente haciéndolo imposible de interpretar. Por lo tanto, el rendimiento relativo se muestra como las diferencias acumuladas de los niveles de CRPS con respecto al benchmark (nuevamente menor nivel es preferido).

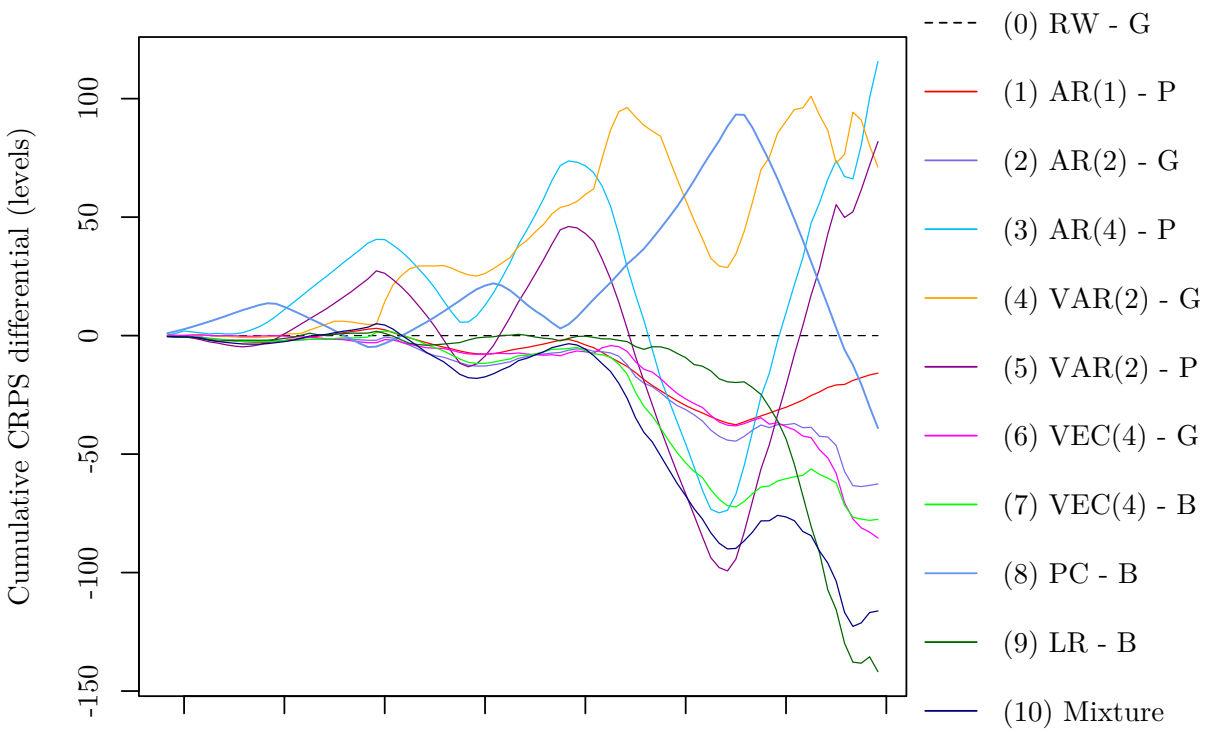


Figura 6: CCRPS en la ventana out-of-sample para $h = 12$

Curiosamente, la mayoría de los modelos tuvieron un rendimiento similar al benchmark hasta 2016-2017, luego el rendimiento de la mayoría de los modelos se desvió significativamente. En general, se puede argumentar que el benchmark falló en capturar la aceleración de la inflación en los periodos del 2016-2019 frente a otros modelos con relaciones a largo plazo como los modelos VEC o el modelo Long-Run que incluían la evolución de los agregados monetarios y salarios como inputs del modelo. Por último, a pesar de que el modelo Long-Run superó al resto de modelos en el horizonte 12, el rendimiento de este modelo solo mejoró notoriamente en los últimos 24 meses. Aunque un modelo como este es una opción a considerar, en la práctica, quizás sea más apropiado buscar modelos con un rendimiento consistentemente mejor en toda la muestra en lugar de algunos períodos específicos en el tiempo. En este caso, el mixture de modelos, a pesar de tener una puntuación CRPS final más baja que el modelo Long-Run, tiene un rendimiento acumulativo consistentemente mejor en toda la muestra con la excepción de los últimos períodos. Este resultado destaca el atractivo de las combinaciones de modelos, ya que pueden tener un rendimiento mejor pero también más estable que un solo modelo.

También se realizó una evaluación de los point forecasts tomando la mediana de la predicción probabilística. La figura (7) muestra el MPE a través de los horizontes. Los resultados indican que tres de los modelos seleccionados (modelos 3, 4 y 5), presentaron un sesgo notable en sus predicciones en horizontes más largos. El resto de los modelos exhibieron un sesgo de menos de ± 1 %.

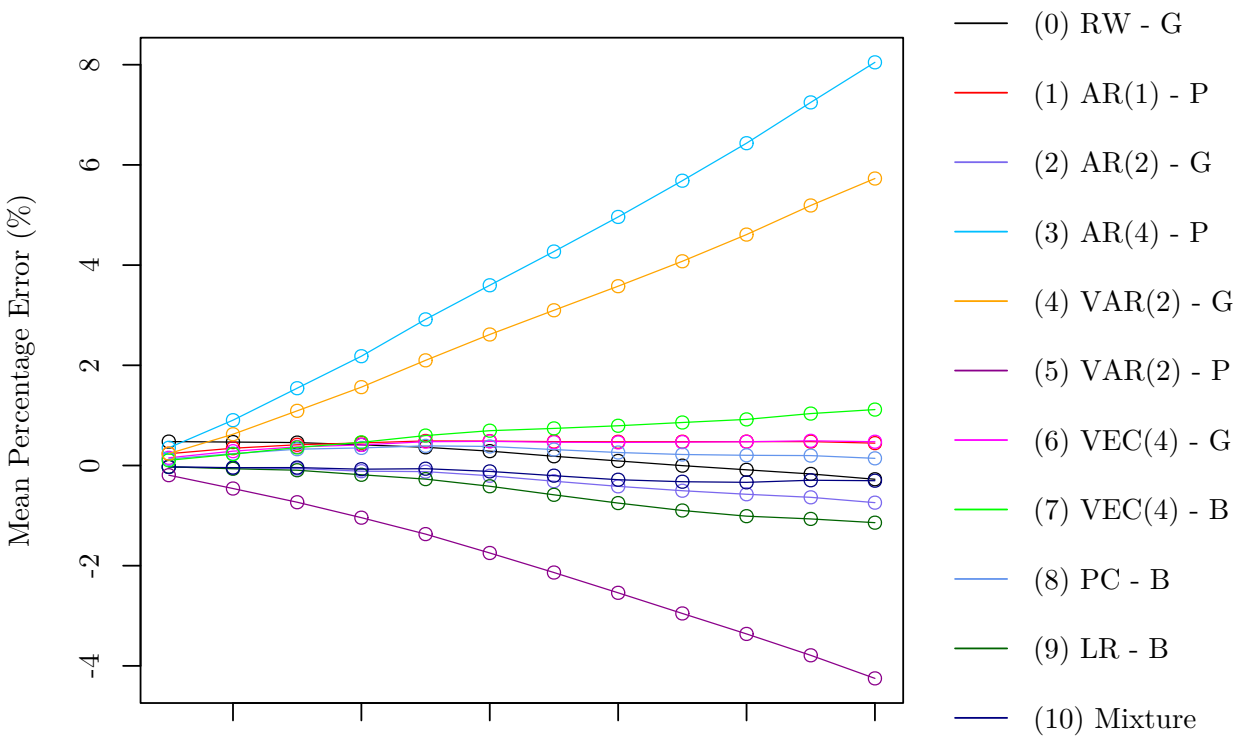


Figura 7: MPE por horizonte

La figura (8) compara el RMSE del point forecast. Esta métrica revela resultados ligeramente diferentes a los del análisis CRPS. Por ejemplo, observe que el desempeño con respecto al benchmark empeoró significativamente para los modelos 3, 4 y 5. Esto está asociado con el hecho de que la mediana de estos modelos exhibió un sesgo notorio en horizontes más largos, sin embargo, los modelos mostraron algunas mejoras en las colas, mejorando la puntuación general de CRPS. A su vez es importante destacar que, con la excepción del modelo de Long-Run, las ganancias relativas del RMSE del resto de los modelos se redujeron notablemente en contraste con los resultados mostrados por el CRPS.

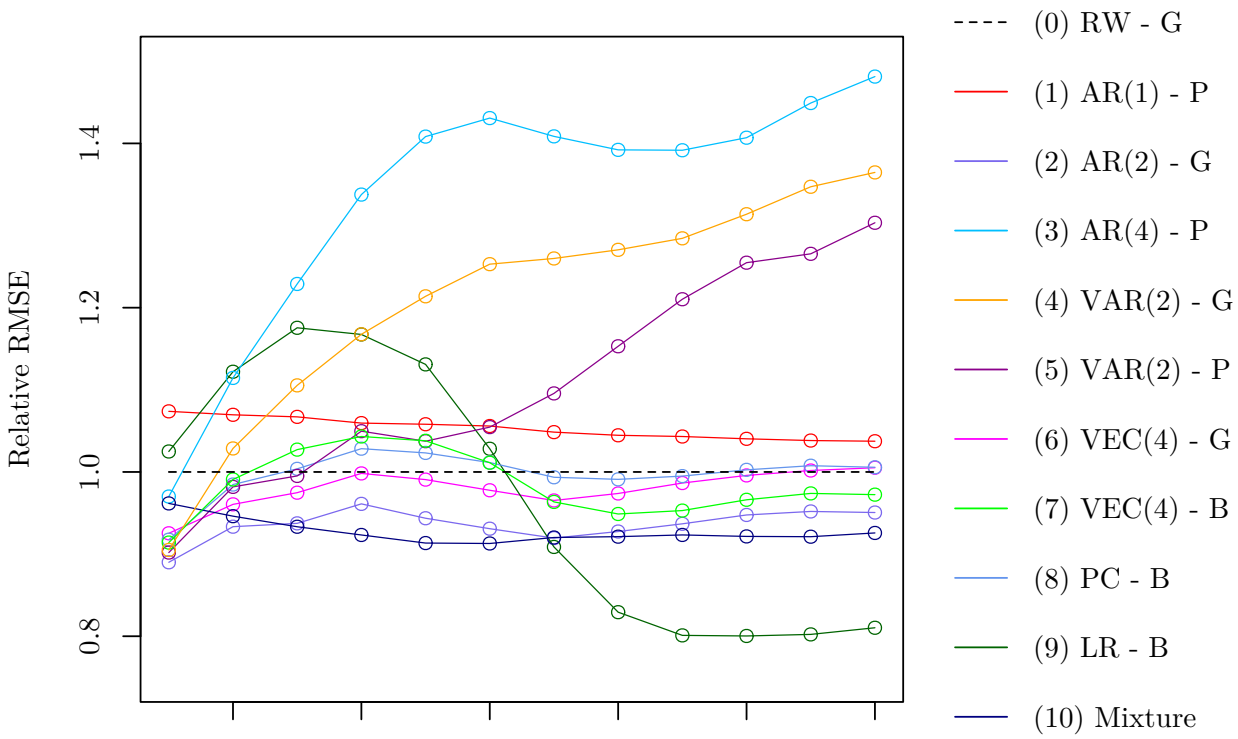


Figura 8: RMSE relativo por horizonte

Aunque esto era algo esperado, ya que el Quantile score mostró un indicio de que los modelos tienden a tener un mejor rendimiento en las colas que en la mediana[17], las implicaciones siguen siendo significativas. Se podría argumentar que, en general, la diferencia de rendimiento entre los modelos multivariados con modelos parsimoniosos, como los modelos AR y de Random Walk, podría no ser tan evidente en los point forecasts, no obstante, existe una oportunidad para explotar modelos multivariados en los forecasts probabilísticos, incluyendo modelos con teoría económica explícita, ya que podrían capturar otras dinámicas integradas que no están presentes en escenarios "normales".

6.2. Resultados del DM test

Se aplicó un DM test para evaluar formalmente el rendimiento predictivo de los modelos probabilísticos. Se eligió probar específicamente la capacidad predictiva del mixture de modelos, el AR(2) y el Random Walk. El modelo univariado seleccionado se eligió porque era el mejor modelo univariado en todos los horizontes. El mixture, por otro lado, no fue superior en todos los horizontes, pero fue seleccionado entre todos los modelos multivariados porque arrojó la mejor puntuación CRPS promediada

[17] Aunque esto solo se comprobó para el horizonte 12, pero no es irrisorio asumir una situación similar para los otros horizontes.

entre los distintos horizontes. Se probó la capacidad predictiva de ambos modelos frente al Random Walk, pero también entre sí. Como se mencionó anteriormente, la prueba solo es válida para aquellos horizontes donde el diferencial del CRPS es estacionario. Se aplicó un test Dickey-Fuller aumentado (ADF) en los diferenciales de CRPS a través de los horizontes (ver el cuadro 2). Los horizontes que fallaron en rechazar la hipótesis de raíz unitaria fueron descartados para el DM test.

Modelo 1 - Modelo 2	h_1	h_2	h_3	h_4	h_5	h_6	h_7	h_8	h_9	h_{10}	h_{11}	h_{12}
AR(2) - RW	0.01*	0.02*	0.01*	0.03*	0.06*	0.11	0.08*	0.06*	0.10*	0.9*	0.16	0.28
Mixture - RW	0.21	0.32	0.06*	0.08*	0.05*	0.05*	0.9*	0.8*	0.11	0.07*	0.09*	0.16
Mixture - AR(2)	0.01*	0.01*	0.12	0.32	0.23	0.09*	0.09*	0.10*	0.8*	0.05*	0.06*	0.05*

Horizontes que exhibieron un valor $p \leq 0{,}1$ son representados por $(*)$.

Cuadro 2: p-values del test de raíz unitaria Dickey-Fuller aumentado (ADF)

Basándose en el cuadro 2, se encontraron los horizontes aceptados y se aplicó el DM test utilizando la siguiente premisa.

Hipótesis Nula: El Modelo 1 y Modelo 2 tienen igual capacidad predictiva.
Hipótesis Alternativa: El Modelo 1 tiene una capacidad predictiva superior al Modelo 2.

Modelo 1 - Modelo 2	h_1	h_2	h_3	h_4	h_5	h_6	h_7	h_8	h_9	h_{10}	h_{11}	h_{12}
AR(2) - RW	0.01*	0.01*	0.02*	0.02*	0.03*	-	0.16	0.19	0.17	0.37	-	-
Mixture - RW	-	-	0.30	0.14	0.07*	0.03*	0.02*	0.02*	-	0.03*	0.01*	-
Mixture - AR(2)	0.27	0.30	-	-	-	0.22	0.16	0.28	0.33	0.09*	0.4*	0.01*

Horizontes que exhibieron un p-value $\leq 0{,}1$ son representados por $(*)$.

Cuadro 3: p-values del test de comparación predictiva de Diebold-Mariano

El cuadro 3 ilustra los p-values del DM test entre el Random Walk contra el modelo univariado y el mixture. El mixture no logra rechazar la hipótesis nula en horizontes más cortos (h_3, h_4), mientras que el modelo AR falla en rechazar la hipótesis en horizontes más largos (h_7, h_8, h_9, h_{10}). Cuando se compararon entre sí, el mixture superó al modelo AR en (h_{10}, h_{11}, h_{12}). Los resultados están en línea

con algunas de las premisas comunes en la literatura de forecasting macroeconómico; 1) El Random Walk o algunos otros tipos de modelos parsimoniosos tienden a ser tan buenos como los modelos multivariados en horizontes más cortos, pero pueden tener un rendimiento inferior en horizontes más largos. 2) El uso de modelos multivariados con mayor nivel de sofisticación puede ser más efectivo para predecir horizontes más largos.

6.3. Resultados de la evaluación PIT

Se realizó una evaluación PIT para el mixture de modelos para evaluar la calibración del modelo. Una vez más, se eligió el mixture entre todos los modelos, ya que tuvo el mejor rendimiento promediado entre todos los horizontes. Dado que el índice de precios al consumidor tiene autocorrelación naturalmente, el análisis se hace en términos de variaciones porcentuales. Siguiendo a (Rossi, 2014), se construyó un histograma y un gráfico ACF de los PITs. La figura (9) no muestra ninguna señal de que el modelo este incorrectamente especificado. Los PITs no exhiben autocorrelación lo cual sugiere independencia, y el histograma revela una distribución relativamente uniforme, esto sugiere que la calibración es correcta.

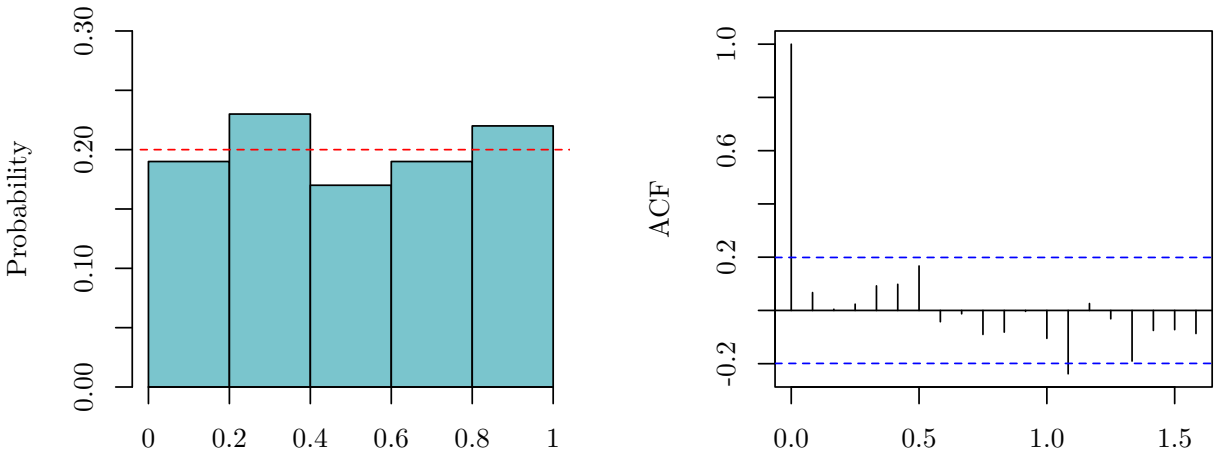


Figura 9: Resultados de la evaluación PIT para predicciones one-step-ahead

Si las predicciones carecieran de calibración, la forma del histograma revelaría la naturaleza del error en la especificación. Por ejemplo, una forma $\bigcup$ es un signo de falta de dispersión, ya que muchas observaciones se consideran demasiado extremas cuando, de hecho, son más comunes en la práctica, lo que sugiere que la densidad predictiva es demasiado angosta. Por el contrario, un exceso de dispersión se refleja en una forma de montaña o $\bigcap$ ya que las distribuciones son demasiado amplias. El sesgo causa inclinaciones o formas triangulares en el histograma hacia un extremo, generalmente una forma de “L” o “J”, dependiendo de la dirección del sesgo.

7. Conclusiones

El documento explora el uso de forecast probabilísticos para predecir el índice de precios al consumidor en la Argentina utilizando una variedad de modelos autorregresivos. Se utilizaron diferentes métricas para evaluar el rendimiento del point forecast, pero también el forecast probabilístico a lo largo de diferentes horizontes. Se utilizó el test Diebold-Mariano en modelos seleccionados para probar la capacidad predictiva. Por último se aplicó la transformada integral de probabilidad para el modelo multivariado escogido y los resultados cualitativos sugieren que el modelo está correctamente especificado.

Los resultados muestran que algunos de los modelos superaron estadísticamente al benchmark en determinados horizontes, pero no hubo un modelo único que superara al benchmark en todos los horizontes. En general, los modelos con estructura (ya sea modelos VEC o modelos con relaciones teóricas explícitas) tuvieron un mejor desempeño.

Un punto central de este análisis es que la diferencia en performance entre los modelos multivariados respecto a los univariados, en general, no fue tan evidente en los point forecasts, pero si en la evaluación de la distribución. Por ejemplo, el rendimiento entre el Random Walk y el mixture de modelos es relativamente similar en la mediana, sin embargo, el mixture es significativamente mejor para capturar los riesgos de cola. Esto sugiere que utilizar modelos multivariados es una estrategia factible para obtener ganancias predictivas en forecast probabilísticos, posiblemente por que su estructura más compleja logra capturar otras dinámicas que no están presentes en escenarios "centrales".

Se usaron mixtures de modelos igualmente ponderados como una forma de explorar combinaciones de modelos y el resultado favorable va en sintonía con la literatura de forecasting. La investigación futura deberá incorporar combinaciones dinámicas como *Dynamic Model Averaging* (DMA) (Koop y Korobilis, 2012). Una linea de investigación posterior debería replicar el ejercicio de evaluación probabilística pero extendiendo la rama de modelos, incluyendo modelos de Machine-Learning, modelos de factores o modelos DSGE que suelen utilizar los bancos centrales. Por último, para el caso particular de la Argentina dado que la serie de IPC tiene un grado de integración de orden dos ($X \sim I(2)$) agregar componentes no lineares (términos cuadráticos o thresholds) podría ser un elemento para incrementar las ganancias predictivas.

8. Apéndice

	Tratamiento Var.	Exp.	EMAE	Tipo de cambio	Salario	MS	Tasa Int.	EEUU CPI	EEUU Tasa int.
(11) AR(1)	Garch (1,1)								
(12) AR(2)	Paramétrico								
(13) AR(3)	Paramétrico								
(14) AR(3)	Garch (1,1)								
(15) AR(4)	Garch (1,1)								
(16) VAR(2)	Bootstrap		X		X		X		
(17) VAR(3)	Paramétrico	X		X		X	X		
(18) VAR(3)	Garch (1,1)	X		X		X	X		
(19) VAR(3)	Bootstrap	X		X		X	X		
(20) VAR(4)	Paramétrico	X	X		X		X		X
(21) VAR(4)	Garch (1,1)	X	X		X		X	X	
(22) VAR(4)	Bootstrap	X	X		X		X	X	
(23) VEC(2)	Paramétrico		X			X			
(24) VEC(2)	Garch (1,1)		X			X			
(25) VEC(3)	Bootstrap			X		X			
(26) VEC(3)	Garch (1,1)			X		X			
(27) VEC(4)	Paramétrico	X		X			X	X	X
(28) VEC(4)	Garch (1,1)	X		X			X	X	X
(29) PC	Paramétrico	X	X	X				X	
(30) Long-Run Alt	Bootstrap		X	X			X	X	X

Cuadro 4: El resto de los modelos y las variables incluidas

Modelo	h_1	h_2	h_3	h_4	h_5	h_6	h_7	h_8	h_9	h_{10}	h_{11}	h_{12}
Random Walk	0.622	1.181	1.745	2.304	2.863	3.399	3.964	4.574	5.193	5.81	6.398	7.30
Modelo 11	0.741	1.651	2.451	3.136	3.806	4.603	5.226	6.002	6.729	7.541	8.744	10.094
Modelo 12	0.749	1.683	2.494	3.208	3.92	4.781	5.473	6.351	7.233	8.18	9.668	11.415
Modelo 13	0.576	1.17	1.828	2.573	3.199	3.766	4.266	4.768	5.289	5.955	6.764	7.66
Modelo 14	0.591	1.233	1.91	2.596	3.213	3.861	4.582	5.427	6.345	7.361	8.392	9.432
Modelo 15	0.665	1.447	2.282	3.117	3.969	4.772	5.609	6.625	7.702	8.659	9.479	10.159
Modelo 16	0.567	1.211	1.871	2.543	3.123	3.803	4.544	5.405	6.382	7.402	8.417	9.412
Modelo 17	0.549	1.144	1.803	2.511	3.209	3.917	4.564	5.311	6.108	6.994	7.916	8.724
Modelo 18	0.628	1.279	1.941	2.641	3.264	4.016	4.779	5.767	6.878	8.033	9.178	10.237
Modelo 19	0.798	1.678	2.371	3.353	4.092	4.975	5.708	6.613	7.66	8.542	9.911	11.395
Modelo 20	0.580	1.212	1.877	2.562	3.162	3.737	4.243	4.805	5.427	6.166	6.977	7.843
Modelo 21	0.694	1.417	2.153	2.974	3.537	4.185	4.724	5.217	5.991	6.904	8.011	9.256
Modelo 22	0.717	1.541	2.398	3.264	4.077	5.003	5.733	6.372	7.367	8.625	10.117	11.832
Modelo 23	0.689	1.479	2.291	3.257	3.988	4.711	5.388	6.132	7.106	8.293	9.454	10.702
Modelo 24	0.739	1.588	2.374	3.423	4.171	4.959	5.673	6.449	7.366	8.467	9.576	10.646
Modelo 25	0.536	1.068	1.66	2.267	2.847	3.452	4.044	4.701	5.427	6.239	7.103	7.909
Modelo 26	0.755	1.599	2.34	3.343	4.343	5.252	6.114	7.201	8.398	9.683	11.247	12.979
Modelo 27	0.944	1.933	2.557	3.215	3.812	4.338	4.969	5.737	6.122	6.942	8.146	9.307
Modelo 28	0.924	1.884	2.594	3.188	3.867	4.531	5.125	5.888	6.657	7.406	8.215	9.184
Modelo 29	0.764	1.613	2.35	3.026	3.861	4.608	5.426	6.527	7.576	8.759	9.975	11.316
Modelo 30	1.013	2.145	2.835	3.479	3.883	4.122	4.599	5.284	6.037	6.306	6.905	8.052

Cuadro 5: CRPS por horizonte para el resto de modelos

Referencias

Agrawal, A., Gans, J., y Goldfarb, A. (2018). *Prediction machines: the simple economics of artificial intelligence*. Harvard Business Press.

Bollerslev, T. (1986). Generalized autoregressive conditional heteroskedasticity. *Journal of econometrics*, *31*(3), 307–327.

Bollerslev, T., Engle, R. F., y Wooldridge, J. M. (1988). A capital asset pricing model with time-varying covariances. *Journal of political Economy*, *96*(1), 116–131.

Brier, G. W. (1950). Verification of forecasts expressed in terms of probability. *Monthly weather review*, *78*(1), 1–3.

Bröcker, J., y Smith, L. A. (2007). Scoring probabilistic forecasts: The importance of being proper. *Weather and Forecasting*, *22*(2), 382–388.

Check, A. J., Nolan, A. K., y Schipper, T. C. (2018). Forecasting gdp: Do revisions matter?

Clemen, R. T. (1989). Combining forecasts: A review and annotated bibliography. *International journal of forecasting*, *5*(4), 559–583.

Cordeiro, C., y Neves, M. (2006). The bootstrap methodology in time series forecasting. *Proceedings of CompStat2006"(J. Black and A. White, Eds.), Springer Verlag*, 1067–1073.

Croushore, D., y Van Norden, S. (2018). Fiscal forecasts at the fomc: Evidence from the greenbooks. *Review of Economics and Statistics*, *100*(5), 933–945.

D'Amato, L., Aguirre, M. G., Garegnani, M. L., Krysa, A., y Libonatti, L. (2018). *Forecasting inflation in argentina: A comparison of different models* (Inf. Téc.). Economic Research Working Papers.

Diebold, F. X., Gunther, T. A., y Tay, A. (1997). *Evaluating density forecasts.* National Bureau of Economic Research Cambridge, Mass., USA.

Diebold, F. X., y Mariano, R. S. (1995). Comparing forecast accuracy. *Journal of Business and*.

Diebold, F. X., y Shin, M. (2017). Assessing point forecast accuracy by stochastic error distance. *Econometric Reviews*, *36*(6-9), 588–598.

Duffie, D., y Pan, J. (1997). An overview of value at risk. *Journal of derivatives*, *4*(3), 7–49.

Efron, B. (1992). Bootstrap methods: another look at the jackknife. En *Breakthroughs in statistics* (pp. 569–593). Springer.

Engle, R. F. (1982). Autoregressive conditional heteroscedasticity with estimates of the variance of united kingdom inflation. *Econometrica: Journal of the econometric society*, 987–1007.

Engle, R. F., y Granger, C. W. (1987). Co-integration and error correction: representation, estimation, and testing. *Econometrica: journal of the Econometric Society*, 251–276.

Garratt, A., Lee, K., Hashem Pesaran, M., y Shin, Y. (2003). A long run structural macroeconometric model of the uk. *The Economic Journal*, *113*(487), 412–455.

Garratt, A., Lee, K., Pesaran, M. H., y Shin, Y. (2003). Forecast uncertainties in macroeconomic modeling: An application to the uk economy. *Journal of the American Statistical Association*, *98*(464), 829–838.

Gneiting, T., y Katzfuss, M. (2014). Probabilistic forecasting. *Annual Review of Statistics and Its Application*, *1*, 125–151.

Gneiting, T., y Raftery, A. E. (2007). Strictly proper scoring rules, prediction, and estimation. *Journal of the American statistical Association*, *102*(477), 359–378.

Granger, C. W. (1981). Some properties of time series data and their use in econometric model specification. *Journal of econometrics*, *16*(1), 121–130.

Hansen, P. R., y Lunde, A. (2005). A forecast comparison of volatility models: does anything beat a garch (1, 1)? *Journal of applied econometrics*, *20*(7), 873–889.

Knüppel, M. (2015). Evaluating the calibration of multi-step-ahead density forecasts using raw moments. *Journal of Business & Economic Statistics*, *33*(2), 270–281.

Koenker, R., y Bassett Jr, G. (1978). Regression quantiles. *Econometrica: journal of the Econometric Society*, 33–50.

Koop, G., y Korobilis, D. (2011). Uk macroeconomic forecasting with many predictors: Which models forecast best and when do they do so? *Economic Modelling*, *28*(5), 2307–2318.

Koop, G., y Korobilis, D. (2012). Forecasting inflation using dynamic model averaging. *International Economic Review*, *53*(3), 867–886.

Lütkepohl, H. (2005). *New introduction to multiple time series analysis*. Springer Science & Business Media.

Manz, C. C., y Sims Jr, H. P. (1980). Self-management as a substitute for leadership: A social learning theory perspective. *Academy of Management review*, *5*(3), 361–367.

Matheson, J. E., y Winkler, R. L. (1976). Scoring rules for continuous probability distributions. *Management science*, *22*(10), 1087–1096.

Riofrío, J., Chang, O., Revelo-Fuelagán, E., y Peluffo-Ordóñez, D. H. (2020). Forecasting the consumer price index (cpi) of ecuador: A comparative study of predictive models. *International Journal on Advanced Science, Engineering and Information Technology*, *10*(3), 1078–1084.

Rossi, B. (2014). Density forecasts in economics, forecasting and policymaking.

Savage, L. J. (1971). Elicitation of personal probabilities and expectations. *Journal of the American Statistical Association*, *66*(336), 783–801.

Schneider, E., Chen, P., y Frohn, J. (2008). A long-run structural macroeconometric model for

germany: An empirical note. *Economics*, *2*(1).

Stenberg, E. (2016). *On the autoregressive conditional heteroskedasticity models.*

Stock, J. H., y Watson, M. W. (2008). Phillips curve inflation forecasts.

Svensson, L. E. (2000). Open-economy inflation targeting. *Journal of international economics*, *50*(1), 155–183.

Winkler, R. L., y Murphy, A. H. (1968). "good" probability assessors. *Journal of Applied Meteorology and Climatology*, *7*(5), 751–758.

Zahara, S., y cols. (2020). Multivariate time series forecasting based cloud computing for consumer price index using deep learning algorithms. En *2020 3rd international seminar on research of information technology and intelligent systems (isriti)* (pp. 338–343).

Zanfei, A., Menapace, A., Brentan, B. M., y Righetti, M. (2022). How does missing data imputation affect the forecasting of urban water demand? *Journal of Water Resources Planning and Management*, *148*(11), 04022060.

Autorizaciones (*En base a su decisión, escoja una opción de cada punto a continuación*)

- **Repositorio Institucional** *(borrar la que no corresponda)***:**

  Autorizo a la Universidad del CEMA a publicar y difundir en el **Repositorio Institucional** de la Universidad de la Biblioteca con fines exclusivamente académicos y didácticos el Trabajo Final de mi autoría.

- **Catálogo en línea** *(borrar la que no corresponda)***:**

  Autorizo a la Universidad del CEMA a publicar y difundir en el **Catálogo en línea** (acceso con usuario y contraseña) de la Biblioteca con fines exclusivamente académicos y didácticos el Trabajo Final de mi autoría.

- **Página web UCEMA** *(borrar la que no corresponda)***:**

  Autorizo a la Universidad del CEMA a publicar y difundir en la **página web de la Universidad** como Trabajo destacado, si el mismo obtuviese la distinción correspondiente, con fines exclusivamente académicos y didácticos el Trabajo Final de mi autoría.

Nombre y apellido: Tomás Marinozzi
DNI: 40010259
Carrera: Maestría en Economía

Firma: