3er Trabajo De Investigación / Informe Ejecutivo: "A new procedure in stock market forecasting based on fuzzy random auto-regression time series model"

Víctor Álvarez
Universidad Técnica Federico Santa María
Vitacura, Chile
victor.alvarezb@usm.cl

Constantino Mavrakis
Universidad Técnica Federico Santa María
Vitacura, Chile
constantino.mavrakis@usm.cl

Matías Otth
Universidad Técnica Federico Santa María
Valparaíso, Chile
matias.otth@sansano.usm.cl

Gabriel Vergara
Universidad Técnica Federico Santa María
San Joaquín, Chile
gabriel.vergara@usm.cl

I. Resumen

El presente informe busca explicar y replicar los resultados obtenidos en el artículo "*A new procedure in stock market forecasting based on fuzzy random auto-regression time series model*" utilizando nuevos datos y mejorando la determinación de los pesos relativos que se utilizan para determinar los outputs a través de un lógica de clasificación del precio de cierre en 2 bandas determinadas por el precio máximo y mínimo en una vela . Este artículo publicado el 10 de febrero del 2018 y fue aceptado por el Journal "*Information Sciences*" perteneciente al cuartil Q1.

La mayoría de las investigaciones con foco en el pronóstico de series financieras se enfocan en preparación de la data, metodologías de pronóstico, evaluación y cuantificación de performance. Sin embargo, la literatura no abunda en la preparación del a data para evitar la aleatoreidad de los precios/retornos, así como la incertidumbre y volatilidad de los precios de las acciones con foco en mejorar la certeza de las predicciones.

El presente articulo, detalla un método de clasificación de preparación de datos, a trravés de una triangulación Fuzzy para después geenrar un mejorado fuzzy random autoregresivo (Fuzzy-AR) modelo para el pronóstico de series no estacionarias (precios).

Los input utilizados son en formato: $[low, high]$ y los precios de cierre por otro lado. Considerando que ladata posee variabildiad y volatilidad, se crean 2 spread *left* y *right*, para luego generar valores esperados y varianzas, de modo de generar intervalos de confianza de la data fuzzy utilziada.

En este contexto de fuzzy input-output, se utiliza el solver simplex para estimar los parámetros del modelo. Encontrando finalmente, que la variabilidad y el ajuste de spread son factores importantes en la preparación de la data para mejorar la precisión del modelo (FUZZY-AR).

Al momento de replicar este paper, se busca mejorar el modelo: Fuzzy Random AR, a través de clasificar el desempeño de la vela, determinando la probabilidad para cada set con la distribución triangular definida anteriormente, clasificando el precio de cierre en dicha distribución.

II. Introducción

En las últimas décadas para el análisis estadístico de series de tiempo a través de modelos, se basaba en regresiones, suaviza miento exponencial, modelos ARIMA-GARCH. Donde la principal debilidad determinada (y abordaba en esta investigación) es que poseen un único input y poseen dificultades para fitear datos no lineales. Lo cual se incrementa considerando que el comportamiento s de los mercados es complejo y se comporta como un sistema dinámico con ruido: no estacionario y definido como una serie caótica. Para solucionar estos problemas es que cobra relevancia los modelos de IA, ya sea, a través de NN, GA o VSM. Sin embargo, estos modelos no son capaces de introducir el concepto de Fuzzy data (en la salida). Es por esto que modelos Fuzzy cobran relevancia en el estudio académico.

El Fuzzy convenciona (FTS) tiene diversas aplicaciones en distintos campos, y ha sido utilziado constantemente para mejorar la performance de predicciones en el mercado accionario.

III. Metodología

La aplicación a utilizar se realizará en función de los retornos logarítimicos de las series de precios a diferencia de los precios nominales dado que la muestra a utilizar incentiva a tener parámetros de autorregresión unitarios por los cambios en la magnitud de los valores en el tiempo.

Para los precios se obtiene el retorno logarítmico como la diferencia logarítmica del cambio de precios mientras que para obtener el retorno *high* y el retorno *low* se obtiene la diferencia del en función del precio pasado versus el valor *high* o *low* de la vela siguiente. En función de esto se trabará con 3 tipos de retonos para una misma serie: el retono usual, el retorno versus su precio máximo alcanzado y verus su precio mínimo alcanzado.

Luego, se utilizará una distribución de membresia triangular simétrica, por lo que se obtiene el retorno medio como el valor promedio entre el retorno high y low, mientras que el spread

del retorno se obtiene como el valor del retorno máximo menos el retorno medio.

III-A. Data

La data a utilizar corresponden a precios de BTCUSDT listados en el mercado de Binance, obtienendo datos desde el 8 de septiembre de 2021 (2:00:00 horas) hasta 29 de octubre de 2022 (16:00:00 horas) a una frecuencia de 2 horas, siendo en total una muestra de 5000 velas.

N°	5000
Max	68524.69 USDT
Min	18067.60 USDT
Media	36869.23 USDT
Std.	13817.45

Cuadro I
RESUMEN SERIE DE COTIZACIONES OBTENIDOS EN LA VENTANA DE TIEMPO ESTABLECIDA

Figura 1. Serie BTC USDT

en cuanto a los retornos de la serie,

N°	4999
Max	0.059246
Min	-0.075362
Media	-0.000154
Std.	0.009946
Skew	-0.086228
Kurtosis	5.995262

Cuadro II
RESUMEN SERIE DE LOG RETORNOS DE COTIZACIONES OBTENIDOS EN LA VENTANA DE TIEMPO ESTABLECIDA

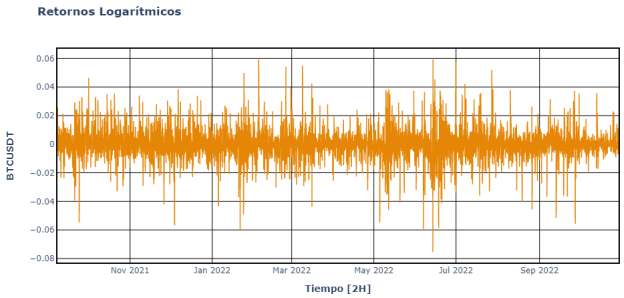

Figura 2. Serie de retornos logarítmicos del periodo de muestra.

III-B. Clasificación fuzzy

Se utiliza la siguiente función de membresia, la cual considera como parámetros. $M =< m, \alpha, \beta >$, que equivalenten a *media*, left y right, respectivamente.

$$\mu(m) = \begin{cases} 0 & si \quad x < m - \alpha \\ 1 - \frac{m-x}{\alpha} & si \quad m - \alpha < x < m \\ 1 & si \quad x = m \\ 1 - \frac{x-m}{\beta} & si \quad m < x < m + \beta \\ 0 & si \quad m + \beta < x \end{cases} \tag{1}$$

Ejemplo distribución triangular:

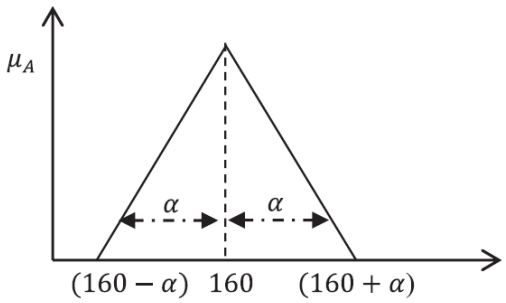

Figura 3. Distribución triangular simétrica

En este caso se consideraría una distribución simétrica, es decir: $(\alpha = \beta)$ con un set asociado: $M =< 160, \alpha >$

III-C. Fuzzy-AR model

- **Modelo AR(p):**

$$y_t = \phi_1 y_{t-1} + \phi_2 y_{t-2} + \ldots + \phi_p y_{t-p} + \epsilon_t \tag{2}$$

- **Output considerando data fuzzy:**

$$Y_t = \bigcup_{i=1}^{n} \left[(Y_{it}^l, Y_{it}^c, Y_{it}^r), P_{it} \right] \tag{3}$$

- **Modelo de regresión Fuzzy:**

$$\tilde{Y}_t = \phi_1 Y_{t-1} + \ldots + \phi_k Y_{t-k} + u \tag{4}$$

- **Modelo FR-AR (2 rezagos):**

$$(Y_t)_T = \left[\theta_1^l, \theta_1^r\right] (Y_{t-1})_T + \left[\theta_2^l, \theta_2^r\right] (Y_{t-2})_T + \left[\varepsilon_t^l, \varepsilon_t^r\right] \tag{5}$$

III-D. Implementación

- **Step 1:**
  Lo primero es realizar la *fuzzificación* de los retornos, para esto , desde la serie de precios se obtienne los retornos logarítmicos : $\{P_t, t \in T\}$ siendo los precios de cierre en el tiempo, luego se definen los precios logarítmicos $p_t = log(P_t)$, de esta manera se tiene los retornos: $r_t = p_t - p_{t-1}$. Además se tiene la serie de precios máixmos y mínimos $H_t, t \in T$ y $\{L_t, t \in T\}$ respectivamente, por lo tanto los retornos respecto al máximo y mínimos se obtienen : $r_t^h = log(H_t) - p_t$ y $r_t^l = log(L_t) - p_t$ respectivamente.
  De este modo se determina la serie de retornos medios y spread:
  $$\begin{aligned} M_t = & \quad \frac{r_t^h - r^{l}{}_t}{2} \\ \alpha_t = & \quad r_t^h - M_t \end{aligned}$$
  de esta manera, el retorno fuzzy $r_t =< M_t, \alpha_t >$ donde M_t corresponde al valor central de la distribución de pertenencia triangular ($\mu(x)$)y α_t es la dispersión a los lados.
- **Step 2:**
  Una vez obtenidos retornos fuzzy se requiere obtener el grado de pertenencia , esto se obtiene desde el retorno realizado versus su posición en la distribución de posibilidad, para ello la función de pertenencia:
  $$\mu_{m,\alpha}(x) = \begin{cases} 1 - \frac{m-x}{\alpha} & m - \alpha < x \leq m \\ 1 - \frac{x-m}{\alpha} & m < x \leq m + \alpha \\ 0 & \text{e.o.c} \end{cases}$$
  así se obtiene el grado de pertenencia de cada retorno como $Pr_{high}(r_t) = \mu_{m_t,\alpha_t}(r_t)$ y $Pr_{low} = 1 - Pr_{high}(r_t)$.
- **Step 3:**
  La autorregresión fuzzy se adapta como:
  $$\begin{aligned} r_t = & \quad \mathbf{1}_{x>0,5}(Pr_{high}(r_{t-1}))\phi_1^r r_{t-1} \\ & + \mathbf{1}_{x>0,5}(Pr_{low}(r_{t-1}))\phi_1^l r_{t-1} \\ & + \mathbf{1}_{x>0,5}(Pr_{high}(r_{t-2}))\phi_2^r r_{t-2} \\ & + \mathbf{1}_{x>0,5}(Pr_{low}(r_{t-2}))\phi_2^l r_{t-2} + e_t \end{aligned}$$
  donde se tienen 2 parámetros por rezago, acompañados de una variable dummy si su grado de pertenencia es mayor al máximo o mínimo.
- **Step 4:**
  Estimar parámetros mediante OLS.

IV. Resultados paper original

En el paper original, el forecasting se realizó a KLSE, el cual hace la distinción entre el Fuzzy-AR de distintos autores y uno propuesto denominado Fuzzy-AR-2 step model.

Se realizaron distintos test y forcast, para compara resultados en distintos sample data y así eliminar el efecto temporal y comparar así solo el performance del los distintos modelos de forecast.

Se puede observar, que el desempeño del modelo Fuzzy propuesto es ampliamente mejor que los distintos modelos econométricos clásicos y que supera también a los anteriores fuzzy's autoregresivos propuestos por la literatura. Independiente del periodo estudiado.

Comparando a través de MSE, se han considerado modelos no difusos y difusos, a saber, SES, DES, ARIMA y FTS. Según las aplicaciones, el modelo FR-AR propuesto con dos tipos de entrada diferentes (Tipo 1 y Tipo 2) supera a otros

Table 12
Evaluation of MSE of KLSE 2009.

Model	MSE	
	Training data (In-sample forecast)	Testing data (out-sample forecast)
ARIMA [4]	138.10	355.97
GARCH [4]	132.24	287.56
SES [5]	274.77	286.86
DES [5]	287.62	299.45
SVM [1]	124.89	221.78
FTS [20]	74.29	91.90
E-FTS [36]	70.77	145.34
FM [1]	72.67	130.97
F-AR [32]	67.02	88.00
FR-AR [32]	62.26	75.87
FR-AR [19]	50.19	50.21
Proposed FR-AR	45.18**	47.20**

** smallest MSE

Table 13
MSE comparisons of KLSE from 2006–2008, and 2016.

Model	MSE (KLSE 2006)		MSE (KLSE 2007)	
	Training	Testing	Training	Training
ARIMA [4]	140.932	234.565	137.643	311.899
GARCH [4]	138.671	210.598	132.765	300.433
SES [5]	99.190	101.120	274.73	278.43
DES [5]	98.012	101.101	287.62	290.64
SVM [1]	129.975	196.658	130.589	287.764
FTS [20]	118.978	122.987	122.632	198.823
E-FTS [36]	115.211	132.245	120.543	207.453
FM [1]	120.764	130.976	129.754	200.634
F-AR [32]	115.108	119.765	119.452	180.765
FR-AR [32]	111.734	117.908	115.723	163.865
FR-AR [19]	105.233	113.551	112.003	143.229
Proposed FR-AR	95.02**	100.06**	100.49**	120.26**

** smallest MSE

Table 13
MSE comparisons of KLSE from 2006–2008, and 2016.

Model	MSE (KLSE 2008)		MSE (KLSE 2016)	
	Training	Testing	Training	Training
ARIMA [4]	136.785	139.533	134.634	137.521
GARCH [4]	131.111	137.231	132.892	135.214
SES [5]	354.577	360.344	116.324	129.869
DES [5]	388.866	392.112	115.734	126.553
SVM [1]	128.451	147.097	140.239	147.198
FTS [20]	124.776	129.761	135.213	138.835
E-FTS [36]	134.876	139.675	140.671	147.642
FM [1]	133.532	137.343	138.832	142.443
F-AR [32]	125.654	129.107	130.872	139.211
FR-AR [32]	122.534	125.635	123.223	133.334
FR-AR [19]	120.111	121.197	120.564	130.113
Proposed FR-AR	115.17**	118.14**	114.02**	125.12**

** smallest MSE

modelos existentes para ambas entradas. Según el hallazgo, este estudio se puede implementar en el manejo de datos de series temporales no estacionarias de varios dominios.

V. Resultados caso 3

Fuzzy LOW/HIGH AR(2):

El primer modelo estimado, es la aplicación de la fuzzyficación de los pares de máximo y mínimo de cada vela con 1 autoregresivo y con 2 autoregresivos, los cuales arrojaron significancia estadística en el primer nivel. Confirmando la hipótesis del paper, donde se afirmaba que los pares máximo y mínimo en su primer rezago y clasificados a través de la lógica explicada en la sección 3 del presente informe; poseen información relevante para realizar la primera estimación del precio de cierre de un activo financiero.

OLS Regression Results

Dep. Variable:	close	R-squared:	0.996
Model:	OLS	Adj. R-squared:	0.996
Method:	Least Squares	F-statistic:	9.375e+04
Date:	Sun, 15 Jan 2023	Prob (F-statistic):	0.00
Time:	19:01:25	Log-Likelihood:	-11081.
No. Observations:	1600	AIC:	2.217e+04
Df Residuals:	1595	BIC:	2.220e+04
Df Model:	4		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	58.8182	38.071	1.545	0.123	-15.855	133.492
Cl_L_L1	0.9737	0.025	38.899	0.000	0.925	1.023
Cl_R_L1	0.9736	0.025	38.880	0.000	0.924	1.023
Cl_L_L2	0.0237	0.025	0.946	0.344	-0.025	0.073
Cl_R_L2	0.0234	0.025	0.933	0.351	-0.026	0.072

Omnibus:	256.262	Durbin-Watson:	1.997
Prob(Omnibus):	0.000	Jarque-Bera (JB):	3044.911
Skew:	-0.339	Prob(JB):	0.00
Kurtosis:	9.724	Cond. No.	1.45e+05

CLOSE AR-1:

A modo de comparación también se realizó el modelo AR-1 de los precios de cierre para tener una comparación con respecto al ajuste dentro de los datos con los que se entrena el modelo y los de forecast.

OLS Regression Results

Dep. Variable:	close	R-squared:	0.996
Model:	OLS	Adj. R-squared:	0.996
Method:	Least Squares	F-statistic:	3.754e+05
Date:	Sun, 15 Jan 2023	Prob (F-statistic):	0.00
Time:	19:02:01	Log-Likelihood:	-11082.
No. Observations:	1600	AIC:	2.217e+04
Df Residuals:	1598	BIC:	2.218e+04
Df Model:	1		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	60.3413	38.018	1.587	0.113	-14.228	134.911
close_L1	0.9971	0.002	612.716	0.000	0.994	1.000

Omnibus:	253.520	Durbin-Watson:	2.042
Prob(Omnibus):	0.000	Jarque-Bera (JB):	3043.721
Skew:	-0.322	Prob(JB):	0.00
Kurtosis:	9.726	Cond. No.	1.44e+05

CLOSE AR-1 y AR-2:

A modo de comparación al igual que el modelo anterior, se defiene el modelo con 2 rezago el cual es significativo estadísticamente solo en el primer nivel.

OLS Regression Results

Dep. Variable:	close	R-squared:	0.996
Model:	OLS	Adj. R-squared:	0.996
Method:	Least Squares	F-statistic:	1.877e+05
Date:	Sun, 15 Jan 2023	Prob (F-statistic):	0.00
Time:	19:05:41	Log-Likelihood:	-11082.
No. Observations:	1600	AIC:	2.217e+04
Df Residuals:	1597	BIC:	2.219e+04
Df Model:	2		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	59.4693	38.030	1.564	0.118	-15.126	134.064
close_L1	0.9737	0.025	38.923	0.000	0.925	1.023
close_L2	0.0234	0.025	0.936	0.350	-0.026	0.072

Omnibus:	255.383	Durbin-Watson:	1.996
Prob(Omnibus):	0.000	Jarque-Bera (JB):	3044.069
Skew:	-0.333	Prob(JB):	0.00
Kurtosis:	9.724	Cond. No.	2.04e+05

Accuracy in sample (MSE):

MES IN SAMPLE	
Fuzzy MSE:	60711.66**
AR2 MSE:	60725.70*
AR1 MSE:	60758.99

Se puede observar que el modelo Fuzzy-AR, es el que incurre en un menor error cuadrático medio al estimar la series de BTCUSD en el periodo estudiado, siendo el con menor desempeño el autoregresivo de primer orden.

Accuracy in sample (MAPE):

MES IN SAMPLE	
Fuzzy MAPE:	161.36*
AR2 MAPE:	161.34**
AR1 MAPE:	161.63

En error absoluto medio, el Fuzzy, corresponde al segundo modelo en cuanto a la métrica.

Accuracy in forecast (MSE): Para el forecasting, AR-

MES IN TEST	
Fuzzy MSE:	18630.84
AR2 MSE:	18550.61**
AR1 MSE:	18556.55*

Fuzzy corresponde al modelo menos efectivo.

Accuracy in forecast (MAPE):

MES IN TEST	
Fuzzy MAPE:	80.65*
AR2 MAPE:	80.61**
AR1 MAPE:	80.96

Finalmente para la métrica MAPE, corresponde al segundo mejor modelo.

VI. Conclusiones

En el paper se afirma que: *El modelo FR-AR se implementó para pronosticar conjuntos de datos reales del mercado de valores. En el caso de datos con una pequeña brecha entre valores altos y bajos, el procedimiento Tipo-1 es más apropiado. Esencialmente, el menor ancho de la posibilidad indica que el modelo se obtiene naturalmente con el procedimiento Tipo-1. En consecuencia, el error puede reducirse. Por otro lado, el procedimiento Tipo 2 es más apropiado cuando existe una gran brecha entre los datos altos y bajos*, lo que se contrasta con los resultados obtenidos en la implementación del modelo.

Como detalla el autor, el Fuzzy-AR-propuesto tipo 1 (replicado en el presente informe), se utiliza principalmente para series con brechas pequeñas, lo cual no se relaciona con la series USDBTC, la cual posee una alta volatilidad de forma intrínseca (agudizada en el periodo estudiado). Es importante recordad que BTCUSD tuvo una caída máxima de cerca del 72 % en el periodo estudiado, por lo que es relevante e considerar que dado la naturaleza de la serie estudiada, el modelo no resulta ser el más eficaz en obtener un performance superior al AR-1 y AR-2.

Se obtienen resultados similares a los AR-1 y AR-2 por lo que no se descarta como un mecanismo con un fuerte potencial para analizar otra clase de activos y con otro sample histórico se podría obtener un mejor desempeño.