# Clustering subsecuencial de series de tiempo: Evidencia de patrones temporales en el tipo de cambio USD/MXN.

**Juan Francisco Muñoz-Elguezábal** [1] **Riemann Ruíz-Cruz** [2]

[1]Msc. Ciencia de Datos - ITESO [2]Departamento de Matemáticas y Físca - ITESO

## Hipótesis - Experimento

**Hipótesis:**
Existe un conjunto de condiciones bajo las cuales la ocurrencia de evento exógeno a una serie de tiempo provoca la presencia de patrones temporales observables en la misma serie.

**Experimento:**
El comunicado de un indicador económico de México o USA es, un evento generador de patrones temporales en la serie de tiempo del tipo de cambio USD/MXN, y evidencia para rechazo de *Hipótesis del Mercado Eficiente = Los precios no son predecibles*. Esto particularmente para el caso de precios intradía.

## Informacion general

- 10 años de información.
- 2010-01-01 al 2020-01-03
- 14.5 Millones de precios.
- 36,000 comunicados de indicadores.
- indicadores económicos de *factset*.
- precios del broker regulado *Oanda*.
- Proyecto con R, Python & LaTeX

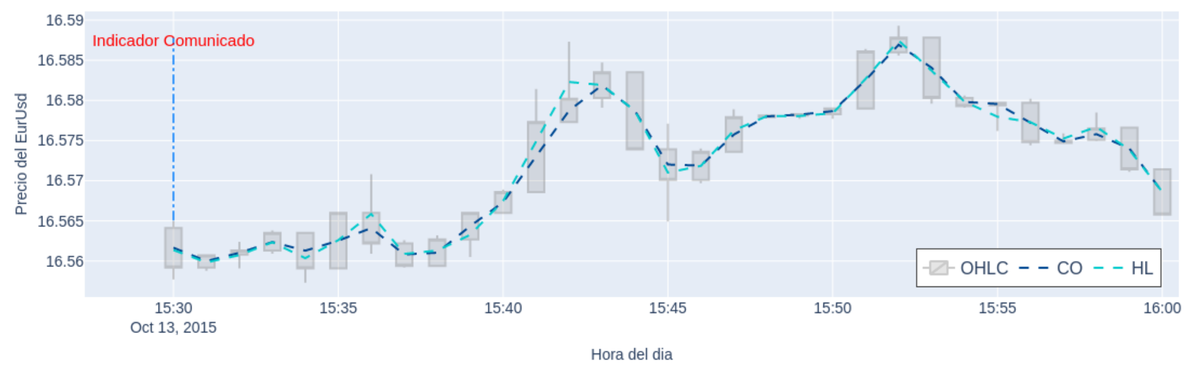
## Definiciones

Sea $\{S_T\}_{T=1}^{n}$, los precios $OHLC_T : \{Open_T, High_T, Low_T, Close_T\}$ de cada $n$ minuto *bursátil* de los últimos 10 años. De los cuales se extraen ventanas de tamaño $m$, de tal manera que $\{OHLC_t\}_{t=1}^{m}$, para $m \in \{10, 20, 30\}$.

- **micro-volatilidades:** $\rightarrow HL_t = (High_t + Low_t)/2$
- **micro-tendencias:** $\rightarrow CO_t = (Close_t + Open_t)/2$

Sea el proceso $\{I_t\}_{t=1}^{k}$ como el comunicado de un indicador macroeconómico que sucede $k$ veces, tal que, con $OHLC_k \{Open_{t=1:m}^{k}, High_{t=1:m}^{k}, Low_{t=1:m}^{k}, Close_{t=1:m}^{k}\}$ se definen las series de tiempo $CO_{t=1:m}^{k}$ y $HL_{1:m}^{k}$ como *motifs* para encontrar en $OHLC_T$

## e.g: Ventana $k = 1$ de $m = 30$ Precios OHLC




## Tipos de indicadores

| categoria | usa | mex | total |
|---|---|---|---|
| actividad economica | 26 | 7 | 33 |
| consumo | 29 | 5 | 34 |
| energia | 4 | 0 | 4 |
| flujos de capital | 5 | 0 | 5 |
| inflacion | 0 | 4 | 4 |
| mercado inmobiliario | 11 | 1 | 12 |
| mercado laboral | 13 | 2 | 15 |
| subasta de bonos | 5 | 0 | 5 |
| tasas de interes | 1 | 1 | 2 |
| total | 94 | 20 | 114 |

## Categoría según resultado

$\{I_t\}_{t=1}^{k} = \{a_t, c_t, p_t\}$ el comunicado de un indicador económico $I$, con información en $t$ de $actual_t$, $consenso_t$, $previo_t$.

$I_t = \{A, B, C, D\} \forall \{I_t\}_{t=1}^{k}$ como la categorización de un comunicado de acuerdo a 4 posibles resultados:

- A: $a_t > c_t > p_t$
- B: $a_t > c_t \leq p_t$
- C: $a_t \leq c_t > p_t$
- D: $a_t \leq c_t \leq p_t$

## e.g. Ocurrencias e invarianza en reacciones según categoría

| indicador | A | B | C | D | T |
|---|---|---|---|---|---|
| 6-Month Bill Auction | 246 | 0 | 0 | 149 | 395 |
| Unemployment Rate | 60 | 4 | 6 | 50 | 120 |

ANOVA $\rightarrow h0$ : varianza constante en reacciones de la misma categoría

| name | esc | obs | anova_hl | anova_co |
|---|---|---|---|---|
| 6-Month Bill Auction | A | 246 | 1 | 1 |
| 6-Month Bill Auction | D | 149 | 1 | 1 |
| Unemployment Rate | A | 60 | 1 | 1 |
| Unemployment Rate | D | 50 | 1 | 1 |

## Medida de similitud entre series de tiempo

Sean $x_t, y_t$ dos series de tiempo de longitud $m$, se define $d(x, y)$ como la distancia euclideana z-normalizada, y $corr(x, y)$ como la correlación entre estas

$$\hat{x} = \frac{x_i - \mu_x}{\sigma_x} \quad , \quad \hat{y} = \frac{y_i - \mu_y}{\sigma_y} \quad , \quad d(x, y) = \sqrt{\sum_{i=1}^{m} (\hat{x}_i - \hat{y}_i)^2}$$

$$corr(x, y) = \frac{\sum_{i=1}^{m} x_i y_i - m\mu_x\mu_y}{m\sigma_x\sigma_y} \quad \rightarrow \quad d(\hat{x}, \hat{y}) = \sqrt{2m(1 - corr(x, y))}$$

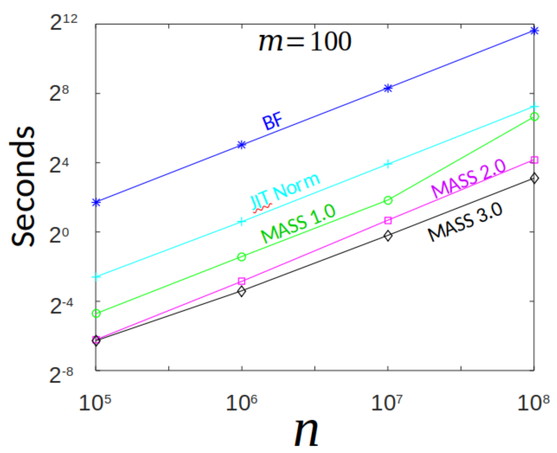
Estadísticas suficientes para cómputo de $d(\hat{x}, \hat{y})$ $\sum_{i=1}^{m} x_i, \sum_{i=1}^{m} y_i, \sum_{i=1}^{m} x_i^2, \sum_{i=1}^{m} y_i^2, \sum_{i=1}^{m} x_i y_i$
$d(\hat{x}, \hat{y}) \implies$ Si se minimiza la distancia se maximiza la correlación.
$d(\hat{x}, \hat{y}) = [0, m_p]$ donde $m_p \in \{0, \mathbb{R}^+\}$.

## Características MASS [1]

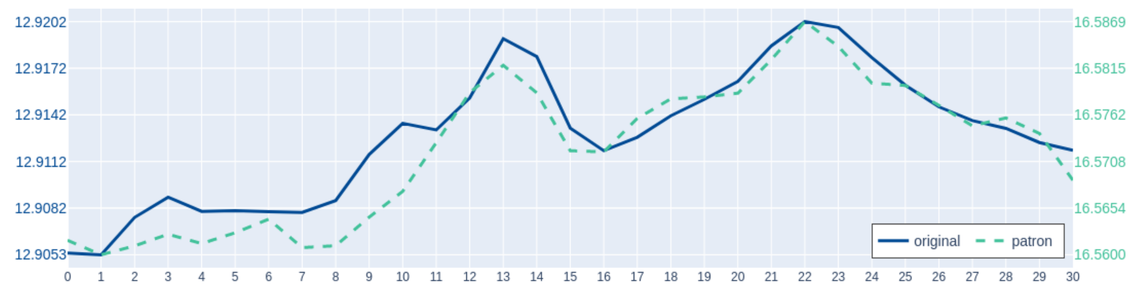
Busca una serie *query* de tamaño $m$ en una serie *principal* de tamaño $n$. Para series de tiempo de cualquier naturaleza.

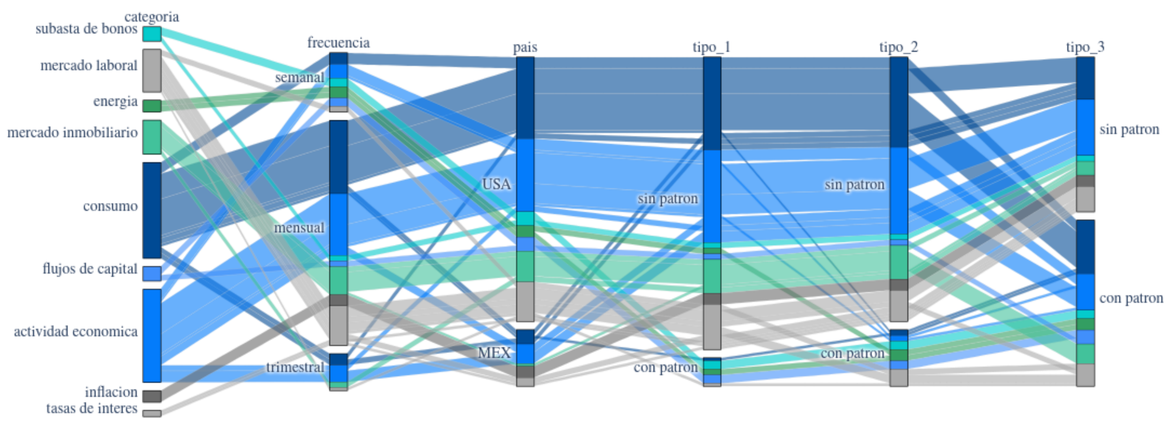
- Implementa una convolución de productos punto móviles con $O(nlogn)$.
- Normalización *just in time*.
- Transformada Rápida de Fourier para convoluciones eficientes.
- Costo computacional no depende de $m$ (libre de efectos de dimensionalidad)

## Complejidad MASS [1]




## e.g. de patron encontrado y resultados generales significativos [2]




## e.g. *Evento disparador de patrones temporales*

e.g. de resultados que arroja búsqueda exhaustiva en paralelo (con 64 núcleos) en 6.5hrs:

| id_esc | tipo_1 | tipo_2 | tipo_3 | total_ind_esc | total_ind |
|---|---|---|---|---|---|
| UnemplntRate_USA_A | 0 | 1 | 15 | 60 | 120 |
| UnemplntRate_USA_D | 0 | 0 | 1 | 50 | 120 |
| 6-Montuction_USA_A | 1 | 2 | 38 | 246 | 395 |
| 6-Montuction_USA_D | 1 | 3 | 10 | 149 | 395 |

tipo_1 : Mismo indicador y categoría  tipo_2 : Mismo indicador diferente categoría  tipo_3 : Diferente indicador y categoría

Query_co = '6-Month Bill Auction - 2014-06-02 15:30 - A' $\rightarrow$ ARIMA (2,1,1)
Patron_co = '6-Month Bill Auction - 2015-10-13 15:30 - A' $\rightarrow$ ARIMA (2,1,2)

- Existen condiciones bajo las cuales se generan patrones temporales
- Modelos predictivos similares para patrones temporales.

[1] Abdullah Mueen, Yan Zhu, Michael Yeh, Kaveh Kamgar, Krishnamurthy Viswanathan, Chetan Kumar Gupta and Eamonn Keogh (2015), The Fastest Similarity Search Algorithm for Time Series Subsequences under Euclidean Distance, URL: http://www.cs.unm.edu/ mueen/FastestSimilaritySearch.html

[2] Chen, J. R. (2007). Making Subsequence TimeSeries Clustering Meaningful, The Encyclopedia of Data Warehousing and Mining (2nd Edition).