Predicción a corto plazo de balanza comercial a partir del índice de tipo de cambio real multilateral y variables productivas con métodos de Machine Learning

Alejandro Watters
Juan Manuel Guzzeti
Cristian Risuelo

UTN - FRBA

***Abstract*— Análisis de correlación a corto plazo entre el tipo de cambio real multilateral e importación, exportación, balanza comercial, producción y consumo con distintos métodos de machine learning.**

Keywords—TCRM, BCRA, producción, exportaciones, importaciones, consumo, acero, competitividad, tipo de cambio. Machine learning.

I. INTRODUCCIÓN

El 16 de febrero de 2017, Federico Sturzenegger, en aquél entonces presidente del Banco Central de la República Argentina (BCRA), hizo un conjunto de declaraciones en las cuales sostenía que la forma correcta de contrastar el tipo de cambio de la economía doméstica con la competitividad era a través del tipo de cambio real multilateral y no del tipo de cambio con el dólar, ya que éste último no tiene lugar en un mundo altamente globalizado en donde el dólar se deprecia frente a otras monedas [1].

Hay evidencia de que la incidencia del tipo de cambio peso-dólar estadounidense es tal que si el tipo de cambio real se apreciara a una tasa de un 1% trimestralmente (manteniendo constante el resto de las variables), el volumen de las exportaciones de manufacturas industriales argentinas descendería en un 6.3% al cabo de 2 años [2].

Por otro lado, el tipo de cambio real multilateral (TCRM) se define como un tipo de cambio ponderado por el comercio entre los países en cuestión. Esto quiere decir que los países que más comercian, en el sector manufacturero, con Argentina, consiguen tener una incidencia en el promedio más grande [3].

Se obtiene a partir de un promedio geométrico ponderado de los tipos de cambio reales bilaterales de los principales socios comerciales del país. Este índice captura las fluctuaciones de las monedas y de los precios respecto de los principales socios comerciales y es, por lo tanto, una de las medidas amplias de competitividad (de tipo precio).

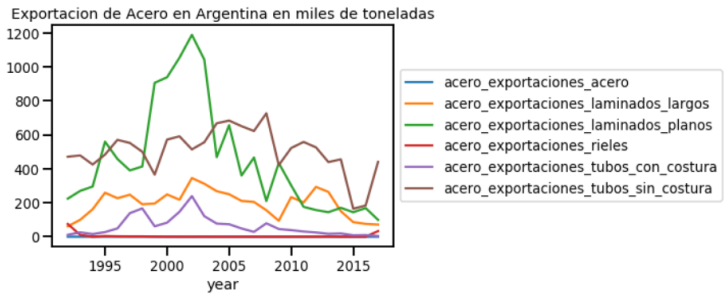
Para nuestro estudio vamos a usar el índice de tipo de cambio real multilateral (ITCRM), ya que el TCRM nominal no nos dice mucho. El ITCRM a utilizar se mide utilizando un año base, en este caso, 2015.

Nuestro objetivo es tener alguna certeza entre las variaciones del ITCRM y los cambios en la producción o en la balanza comercial (exportaciones menos importaciones) utilizando métodos de Machine Learning.

II. ANÁLISIS EXPLORATORIO DE DATOS

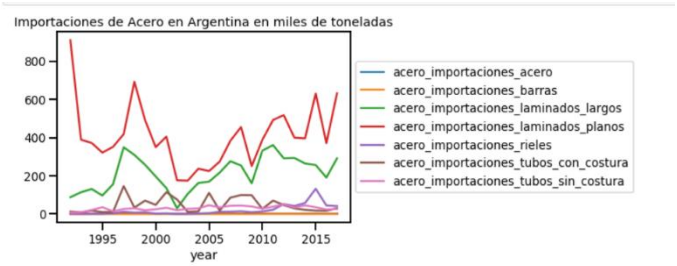
A. *Exportaciones*

Las exportaciones están divididas en varios segmentos: Laminados largos, laminados planos, rieles, tubos con costura y tubos sin costura. Para contabilizar el total sólo se tiene en cuenta el peso en toneladas. Los datos se encuentran registrados de manera trimestral desde 1992 hasta 2019.



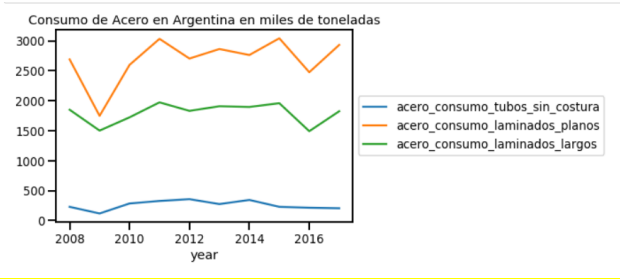
B. *Importaciones*

Las importaciones se encuentran registradas temporalmente de igual manera que las exportaciones, pero difieren en que aquí encontramos una nueva categoría “barras”, las cuales no segregamos al agrupar.



C. *Consumo*

El consumo se trata de un data set más corto ya que solo se conforma de tubos sin costura, laminados planos y laminados largos y se registra desde el año 2008 al 2019.



D. Producción

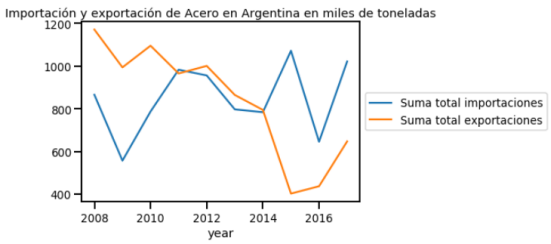
La producción se cuenta como las toneladas de acero fundido que se consiguen en un período y sus datos van del año 1993 a 2019.

E. ITCRM

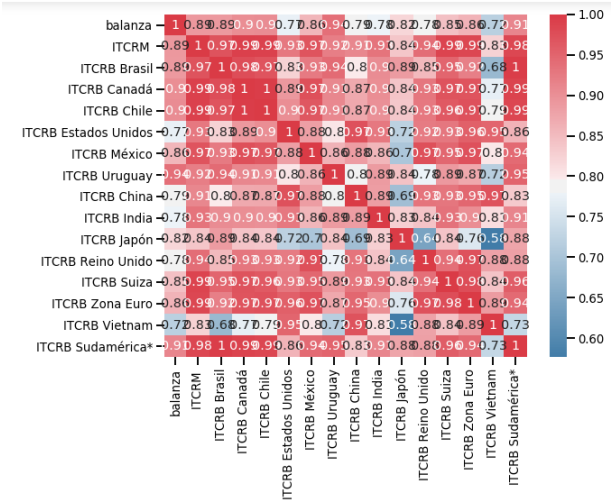
Dicho data set contiene el ITCRB de argentina con sus mayores socios comerciales desde 1997 hasta 2020 de manera diaria. Para poder trabajar en igualdad de condiciones con el resto de los datos, debemos tomar los datos de estos de manera trimestral, por lo que debimos incluir varias líneas de código para pre-procesar este set de datos.

F. Correlaciones

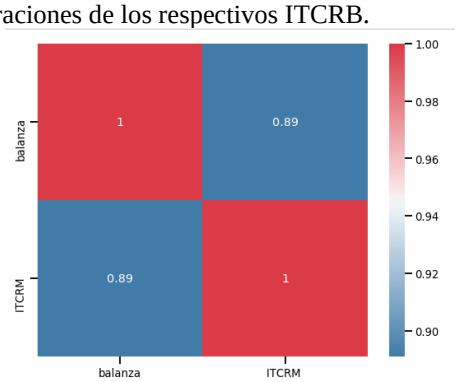
En primer lugar debemos obtener el valor de la balanza comercial, para ello necesitamos hacer la diferencia entre exportaciones e importaciones. La magnitud de medida de estas últimas es el peso en toneladas comercializado en el mismo período de tiempo (un trimestre).



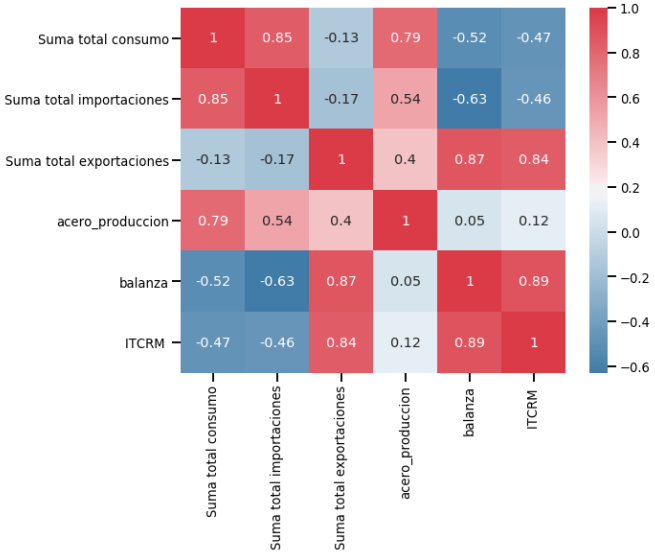
Teniendo todos los data set listos, en esta etapa queremos evidenciar la posible relación entre nuestras variables, es así como procedemos a confeccionar una matriz de correlación entre la balanza comercial y el ITCRM bilateral (ITCRB) con los países que comercian con Argentina que sean más representativos.



Luego de ver el resultado de forma bilateral, pasamos a aislar el ITCRM multilateral, que es aquel que tiene en cuenta todas las ponderaciones de los respectivos ITCRB.



Finalmente procedemos a realizar correlación entre las variables de producción anteriormente explicadas junto al ITCRM



III. ANÁLISIS DE REGRESIÓN

A. *Preparación de data set*

Teniendo nuestro data set con todas las features deseadas debemos realizar unos pasos previos para implementar los modelos de Machine Learning.

En primer lugar, separamos el dataset en valores X e Y donde, como dijimos, Y será el valor de la balanza comercial y seguidamente realizaremos la separación de set de prueba y entrenamiento.

Luego de ello pasaremos a escalar los datos con el algoritmo MinMax Scaler y utilizaremos el algoritmo de Principal Component Analysis [4] buscando con el mismo, reducir la dimensionalidad de nuestros datos, disminuyendo su complejidad pero reteniendo la variabilidad característica del data set original, para optimizar el futuro procesamiento de los datos en cada uno de los modelos de regresión.

Dicho esto, como ultima paso, generaremos features polinómicas de 4to grado para utilizarlos en los modelos, de manera de que podamos ver si al aumentar la dimensión de nuestros datos obtendremos un mejor score final,
Cabe mencionar que todos los modelos serán entrenados tanto con las variables lineares como con las variables polinómicas.

B. *Regresión lineal*

Dado nuestros datos empezamos a analizar por el modelo lineal, siendo constituido éste por aquella recta que minimiza la suma de los residuos al cuadrado.
Es decir, queremos minimizar

$$\mathrm{RSS}(w) = (\mathbf{y} - \mathbf{X}w)^T (\mathbf{y} - \mathbf{X}w)$$

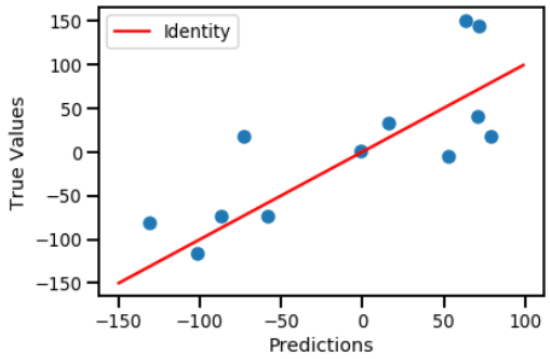
La regresión lineal está determinada por parámetros **w**, que caracterizan a cada feature. Por lo tanto, el procedimiento de optimización será

$$\frac{\partial \mathrm{RSS}}{\partial w} = 0$$

$$\hat{w} = (\mathbf{X}\mathbf{X}^T)^{-1}\mathbf{X}^T\mathbf{y}$$

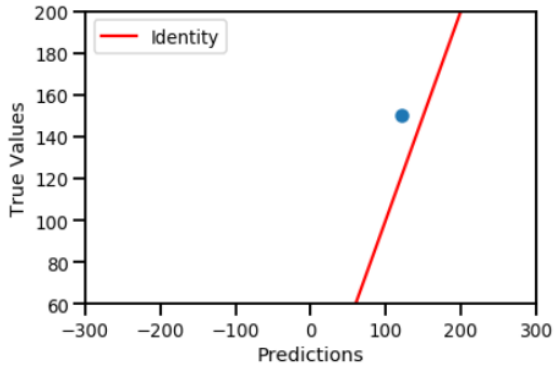
Obteniendo los parámetros procedemos a construir nuestra recta, comparando así las predicciones con los valores reales.

```
R2 score: 0.581110
MAE: 2696.886650
MSE: 42.529977
```



Regresión lineal polinómica

```
R2 score: -0.337681
MAE: 8612.233667
MSE: 71.248828
```



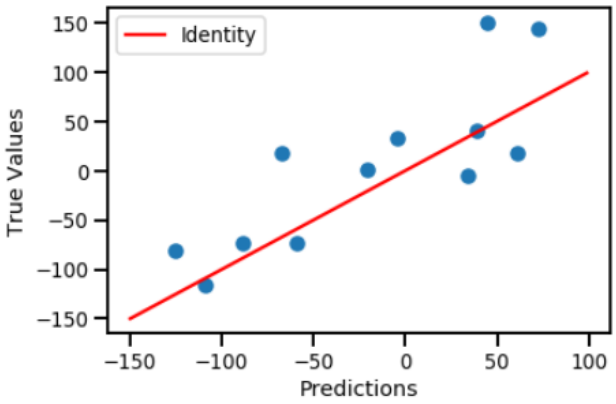
C. *Regresión lineal Ridge*

La regresión Ridge [5] impone una penalización a los parámetros **w** haciendo que estos tiendan a cero en caso de que no sean tan importantes.
Para esto se utiliza la norma L2 del vector **w**

$$\|\mathbf{w}\|_2 = (|w_1|^2 + |w_2|^2 + \cdots + |w_d|^2)^{1/2}$$

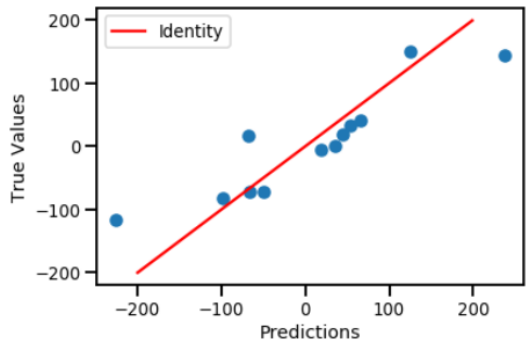
Procedemos ahora a averiguar cuales son los mejores hiperparametros para este estimador, utilizando el algoritmo GridSearch, obteniendo un lambda=1 como mejor score.

A continuación pasamos a hacer predicciones con los parámetros recién obtenidos.



D. Regresión polinómica de Ridge

```
R2 score: 0.571875
MAE: 2756.345177
MSE: 41.058552
```



E. SVR lineal

F.

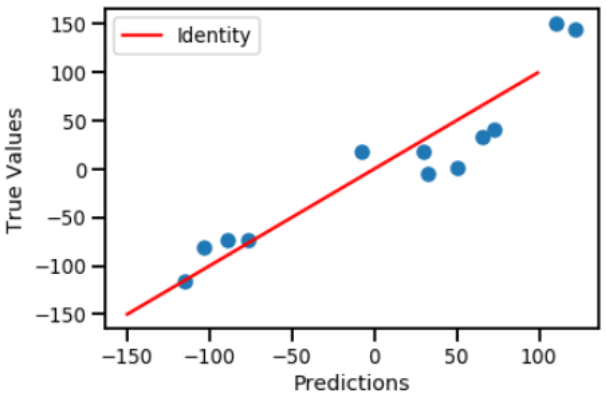
Support Vector Regression [6] es un método que construye una función lineal minimizando el margen entre ella y las muestras. Su forma de trabajo consiste en determinar un margen/radio (epsilon) como función de costo y trata de que todas las muestras estén dentro del margen.

El hiper-parámetro es una función C que penaliza muestras fuera del radio.

$$C\sum_{n=1}^{N}\xi_n + 1/2\,\|w\|^2$$

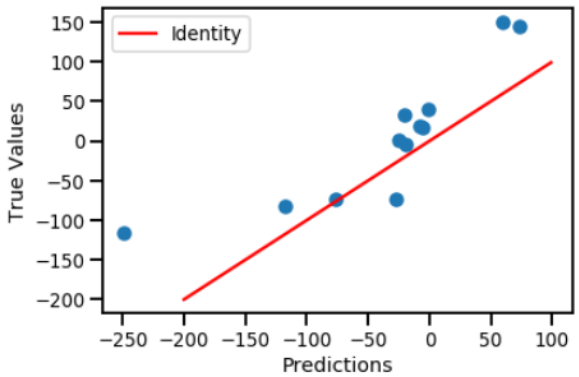
Luego de evaluar la mejor combinación de hiperparametros para el estimador SVR utilizando el algoritmo GridSearch procedemos a entrenar nuestros datos obteniendo la siguiente predicción. Los hiperparametros seleccionados fueron, 'C': 1500, 'epsilon': 0.001, 'gamma': 1.

```
R2 score: 0.876849
MAE: 792.868615
MSE: 24.364338
```



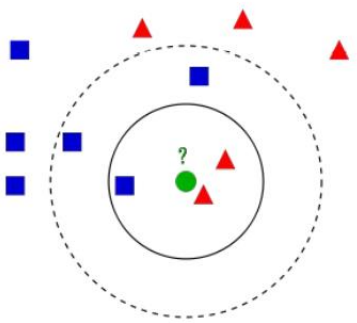
G. SVR polinómico

```
R2 score: 0.469743
MAE: 3413.890079
MSE: 46.987019
```



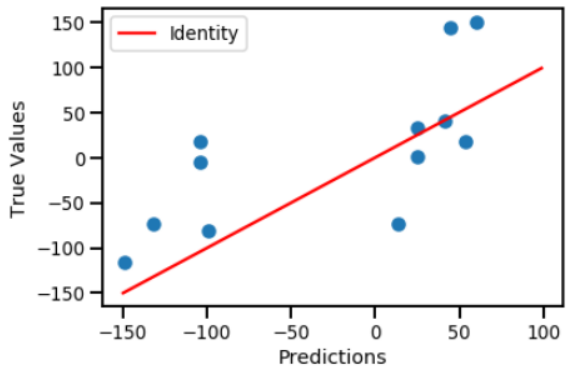
H. KNN lineal

Para realizar KNN en la parte del data set disponible como entrenamiento se debe determinan los K vecinos más cercanos por distancia euclídea. El Y a predecir se determina por la interpolación de los Y dato en los K vecinos.



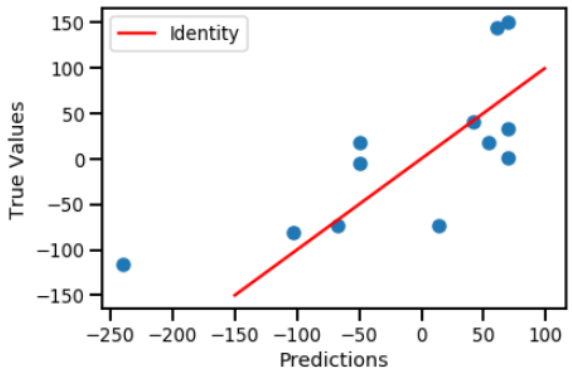
Luego de determinar el valor de K que nos de un mejor score utilizando nuevamente Gridsearch, entrenamos los datos con el modelo para obtener predicciones y resultando.

```
R2 score: 0.257296
MAE: 4781.666042
MSE: 56.454167
```



I. KNN *polinómico*

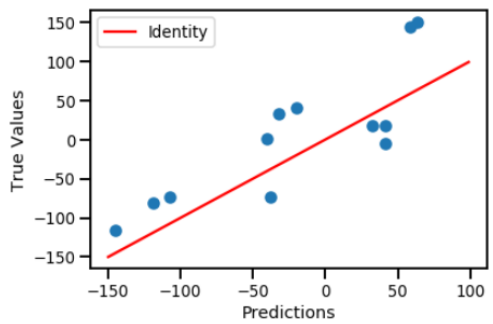
```
R2 score: 0.339980
MAE: 4249.325625
MSE: 54.854167
```



J. Random Forest lineal

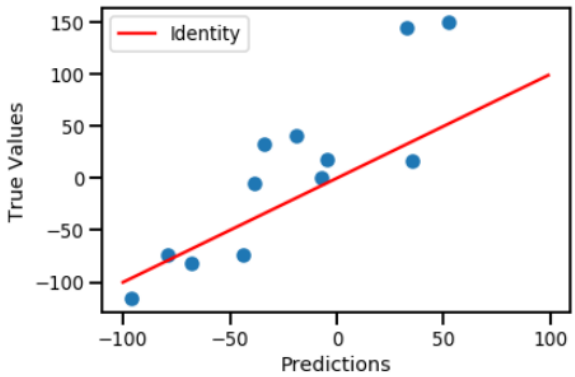
Random Forest [7] es un metaestimador que se ajusta a varios clasificadores de árboles de decisión en varias submuestras del conjunto de datos y usa promedios para mejorar la precisión predictiva y controlar el sobreajuste.
Al igual que en los algoritmos anteriores luego de determinar los mejores valores del parámetro lambda, obtenemos el siguiente modelo de predicciones

```
R2 score: 0.585106
MAE: 2671.160079
MSE: 46.635611
```



K. Random Forest polinómico

```
R2 score: 0.565675
MAE: 2796.262023
MSE: 40.574917
```



IV. CONCLUSIONES

A. Conclusiones sobre el análisis exploratorio de datos.

En primer lugar, en el análisis exploratorio de datos pudimos ver que existe una fuerte correlación entre ITCRM y balanza comercial. Esto está acorde con la teoría económica, donde un aumento del ITCRM implica un aumento en la competitividad, ya que los costos medidos en pesos de los exportadores se licúan, mientras que el precio de venta fijado en el mercado internacional sigue siendo el mismo, ya que se mide en moneda extranjera. Este aumento de la competitividad aumenta el ITCRM y aumenta la balanza comercial.
Es decir que, a corto plazo, existe una correlación entre ITCRM y balanza comercial.
Nuestros métodos de machine learning también nos indican que existe una correlación entre exportaciones e ITCRM, lo cual es lógico, y esto debe ser, en primer lugar, el origen de correlación de ITCRM con balanza comercial. Ya que, como definimos

$$Balanza\ comercial = Exportaciones - Importaciones$$

Por otro lado, vemos que no podemos afirmar que exista correlación entre el nivel de producción y el ITCRM, lo cual suena razonable debido a que el ITCRM es una medida de competitividad externa, y, de existir trabas para la producción en el mercado interno, estas no pueden ser imputables al mercado cambiario, ya que son otro tipo de variables reales las que afectan la producción de un país, como puede ser el esquema impositivo, sistema legislativo y regulatorio y el mercado de capitales existente, por nombrar algunos.

B. Conclusiones del análisis de regresión

El resumen de nuestros procedimientos se expresa en el siguiente cuadro.

RESUMEN DE RESULTADOS				
Model	**Features**	**R2**	**MSE**	**MAE**
Linear	Lineal	0.581	2.696.887	42.530
Linear	Poly	-0.338	8.612.234	71.249
Ridge	Lineal	0.599	2.582.574	40.321
Ridge	Poly	0.572	2.756.345	41.059
SVR	Linear	0.877	792.869	24.364

SVR	Poly	0.470	3.413.890	46.987
KNN	Linear	0.257	4.781.666	56.454
KNN	Poly	0.340	4.249.326	54.854
Rand Forest	Linear	0.585	2.671.160	46.636
Rand Forest	Poly	0.566	2.796.262	40.575

Con claridad se ve que el mejor modelo de regresión para nuestro caso es el SVR lineal, ya que destaca con un buen valor de R2 y una considerable diferencia con respecto a los MSE y MAE del resto de modelos.

Esto tiene lógica teniendo en cuenta el funcionamiento característico del SVR que hemos descripto en el apartado correspondiente, al buscar este un margen optimo en el cual se vean incluidos la mayoría de las muestras y penalizando a aquellos que no lo hagan, tiene sentido el alcanzar un modelo aceptable.

Por otro lado, vemos que todos los modelos poseen altos valores de MSE, esto también tiene sentido, ya que es un valor sensible ante malas predicciones y está tratando con un dataset con datos con alta variabilidad.

Otra conclusión que podemos obtener de este análisis es que el agregado de variables polinómicas no mejoro el rendimiento de los modelos, sino que, por lo contrario, el aumento de dimensionalidad genero mayor error y menor score en prácticamente todos los casos.

[1] Federico Sturzenegger: “Se tiende a poner un énfasis particular en la relación peso-dólar. Sin embargo, nosotros encontramos que esta relación, desde el punto de vista de nuestro comercio exterior, es bastante poco relevante. Más aún en momentos como este, en que el dólar cambia de valor bastante significativamente respecto del resto de las monedas en conjunto. En este mundo globalizado, el tipo de cambio que hay que mirar es el tipo de cambio multilateral, una suerte de combo que toma en cuenta todas las monedas relevantes para nuestro comercio. O mejor aún, el tipo de cambio real multilateral, que corrige por la evolución de los precios domésticos”. Febrero, 2017.

[2] Daniel Berrettoni, “Exportaciones y tipo de cambio real: el caso de las manufacturas industriales argentinas”.

[3] Ponderado por el comercio de manufacturas (promedios móviles 12 meses. Ponderaciones actuales: Brasil: 30%; Estados Unidos: 13%; China: 15%; Zona del euro: 20%) Fuente: BCRA, INDEC, Direcciones de estadística de la Provincia de San Luis y de la Ciudad Aut. de Buenos Aires, Thomson Reuters y REM-BCRA

[4] Wold, S., Esbensen, K., & Geladi, P. (1987). Principal component analysis. Chemometrics and intelligent laboratory systems, 2(1-3), 37-52.

[5] Hoerl, A. E., & Kennard, R. W. (1970). Ridge regression: Biased estimation for nonorthogonal problems. Technometrics, 12(1), 55-67.

[6] Drucker, H., Burges, C. J., Kaufman, L., Smola, A. J., & Vapnik, V. (1997). Support vector regression machines. In Advances in neural information processing systems (pp. 155-161).

[7] Liaw, A., & Wiener, M. (2002). Classification and regression by randomForest. R news, 2(3), 18-22.