Pontificia Universidad Católica de Chile
Escuela de Ingeniería
IIC3695 Tópicos Avanzados en Inteligencia de Máquina 2019-1

Entrega 1 Proyecto Semestral

5 de julio de 2019

Raimundo Herrera - Manuel Vial

1. Introducción

Existe un proyecto en ejecución de crianza de esturiones en Parral, región del Maule. Estos peces viven en piscinas que operan con un sistema llamado de recirculación, que permite reutilizar hasta un 95 % del agua luego de una limpieza de la misma con biofiltros.

Para el correcto crecimiento y desarrollo de los peces, se requiere mantener el agua en buenas condiciones, razón por la cual, ésta debe ser monitoreada constantemente. Para ello, se miden una serie de parámetros que indican de forma directa o indirecta si las condiciones son óptimas para los peces.

Algunos de estos parámetros deben mantenerse dentro de rango a toda costa, pues en caso contrario, la mortalidad puede ascender al orden de miles de peces. Tendría gran valor poder anticipar estos riesgos o mejor aún, evitarlos.

Luego de 5 años de crianza, se está analizando la posibilidad de realizar el monitoreo de parámetros de forma automática en vez de manual como se hace actualmente. Esto implica la compra y mantención de una serie de sensores de alto costo. Por ello, el estudio de relaciones entre parámetros se hace tentador para evitar gastos innecesarios.

Durante estos 5 años de operación, se han tomado datos de 6 parámetros, en todas las piscinas, tanto a la entrada del filtro biológico presente en las mismas, como en la salida, ambas mediciones tomadas una vez al día. Hay datos faltantes como se explicará más adelante y es esperable que también existan errores en las mediciones anotadas, pues se hacen de forma manual.

En la tabla [1] incluida en los anexos, se muestran los principales parámetros medidos actualmente con sus rangos y valores óptimos. Con esos datos a disposición, uno de los resultados más valiosos que puede obtenerse, es la estimación de la concentración de nitrito en el agua, puesto que el sensor que lo mide es el más costoso.

En este trabajo se procede realizando una estimación de los parámetros que rigen la serie de tiempo asociada a la concentración de nitrito. Para ello, en primer lugar se realiza un

análisis exploratorio y visual de los datos para tener una sensibilidad sobre el comportamiento de los mismos, las medidas de dispersión de las distintas variables, las correlaciones y autocorrelaciones de la mismas, junto con observar cada parámetro como una serie de tiempo, para ayudar visualmente a comprender el comportamiento de las curvas.

En segundo lugar, se trabaja con un modelo autorregresivo, particularmente con el llamado ARMA (*Autoregressive Moving Average Models*), que intenta aprovechar la información pasada para calcular el valor presente de un parámetro. Sus coeficientes se calcularán con el método de Metrópolis-Hastings.

2. Marco Teórico

Luego de un análisis acabado de los datos fue posible constatar que usarlos como matriz de diseño no era un opción inteligente pues la autocorrelación es bastante alta y la naturaleza de los datos es altamente dependiente de la temporalidad. Así, lo más lógico es usar métodos que consideren esta relación intertemporal como las series de tiempo. Por ello, se entró a la investigación de modelos tipo ARMA o derivados.

Estos modelos intentan describir la dependencia de una variable buscada con las demás y consigo misma en tiempos pasados. En la mayoría de estos modelos se asume linealidad porque mantiene la simplicidad de las operaciones y no pierde mucho en capacidad de descripción de las realidades que busca predecir.

En términos de ecuaciones, se puede escribir el modelo ARMA como sigue:

$$y_t = \sum_{i=1}^{p} a_i y_{t-i} + \sum_{i=1}^{q} \sum_{j=1}^{M} c_i^j u_{t-i}^j + \epsilon_t$$

En este modelo, se reciben entradas u que, junto a los valores anteriores de y y el error ϵ definen cuál va a ser el nuevo valor para y. Las entradas u corresponden a aquellos parámetros que se consideran como causas directas del valor a predecir y. Por otro lado, p y q expresan cuántas muestras anteriores se toman del propio y y de las entradas, respectivamente. Los subíndices indican temporalidad (en general, discreta), y los supraíndices indican a qué entrada corresponde, pues puede haber más de una. Los coeficientes a y c son los que permiten ajustar el modelo. M corresponde a la cantidad de parámetros que se utilizarán como entrada. En la siguiente sección se explica con más detalle el modelo ARMA.

Los modelos de este tipo son ampliamente usados en la literatura y la industria porque permiten explicar bastante bien la realidad en muchos casos, manteniendo una forma matemática simple y fácil de manejar. Así, existe mucha investigación que relaciona estos modelos con distintos rubros y ocupa técnicas de distintas ramas del conocimiento [1]. La

búsqueda de los coeficientes usando *maximum likelihood* se ocupa desde hace más de 10 años [2].

En este proyecto, se busca predecir el nivel de nitritos. Para ello, como no conviene aumentar demasiado la cantidad de parámetros, se consideran algunos como entrada u y se supone que los demás generan un efecto poco considerable por lo que no se distinguen del ruido o error. La elección de cuáles parámetros se consideran y cuáles no es crucial para los resultados del modelo, por lo que en las secciones venideras se discutirá sobre las distintas elecciones, y su influencia en el rendimiento de la predicción. El modelo asume que existe ruido *gaussiano*.

Se busca encontrar los coeficientes de a_i y c_i^j para cada j y cada i. La forma en este caso, será la maximización de la *likelihood*. La literatura muestra que métodos como Gibbs-Sampling o su versión general, Metrópolis-Hastings pueden ser ocupados para esto [3]. En este caso, se usó Metrópolis-Hastings para obtener los coeficientes, pero se mantiene abierta la posibilidad de probar con Gibbs-Sampling para comparar resultados.

Un elemento a analizar es la estructura de los datos faltantes. Existen 3 mecanismos que la explican, *Missing Not At Random (MNAR)*, *Missing At Random (MAR)* y *Missing Completely At Random (MCAR)* [4]. Para el conjunto de datos utilizado, se considera que la información faltante se rige bajo *MAR*. Lo anterior, si bien no se puede probar sin acceso a la información faltante, se deduce a partir de que la falta de datos no se relaciona con el valor de los datos ausentes, sino que con el valor de otras variables observadas [4]. Según la información entregada por la planta, existen situaciones relacionadas con la operación que determinan que ciertos días en específico no se realicen mediciones. De este modo, como sí existe información observable que da cuenta de la falta de datos, se descarta que sea *MCAR*, y puesto que no depende de los valores de los datos que faltan sino de otras variables observadas, se descarta que sea *MNAR* [4].

Cuando se concluye que la información es *MAR* o *MCAR* se puede proceder ignorándola para el análisis posterior. No obstante, si bien es un análisis relevante de hacer, se ha mostrado que si la suposición de *MAR* es incorrecta, dado que no se puede probar, esto no distorsiona significativamente las estimaciones [4], por lo que asumirlo no implica un riesgo mayor para el modelo.

3. Metodología

Como se mencionó anteriormente, la primera parte del trabajo desarrollado consistió en un análisis exploratorio de los datos, con el objetivo de entender el comportamiento de los mismos.

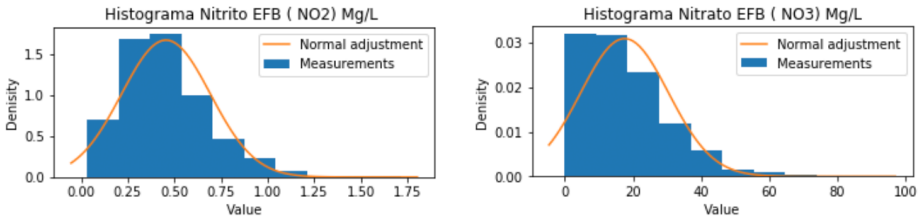

Figura 1: Histograma para mediciones de nitrito y nitrato

3.1. Visualizaciones

En primer lugar, para observar si existen distribuciones que rigen a cada parámetro, se dispusieron histogramas, junto con una distribución normal de acuerdo a la media y desviación estándar de cada uno a modo de referencia. En la figura [1] se muestran dos de los histogramas más importantes -los más correlacionados- el de nitrito y el de nitrato, junto con el ajuste normal. Como se observa, los datos están lejos de distribuir de acuerdo a una distribución normal. Si bien muestran ciertas similitudes en su forma, la correlación, como se mostrará más adelante no es alta.

En segundo lugar, se visualizaron los parámetros como series de tiempo, lo que es de especial interés dado el modelo a utilizar. Las visualizaciones de dos de las más notables se muestran en los anexos [5], sin embargo, como no aportaban mucho, se optó por mostrar series de acuerdo a promedios de las mediciones en un mes, como se ve en la imagen [2], además de las por día, para observar si existían comportamientos a una escala temporal mayor que no se apreciaban con el ruido diario. Si bien esta visualización otorga cierta información respecto a *peaks* compartidos, no es suficiente para argumentar una similitud directa.

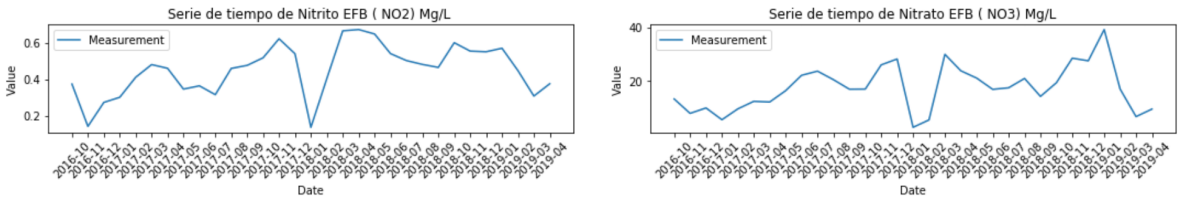

Figura 2: Series de tiempo promediadas por mes para las mediciones de nitrito y nitrato

3.2. Correlaciones

En el análisis de la correlación se encontraron parámetros relacionados entre sí, que confirmaron algunas hipótesis sobre el comportamiento de ciertos compuestos químicos. Por ejemplo, por su composición, el nitrito y nitrato deben estar íntimamente relacionados.

En la matriz de correlaciones mostrada en los anexos [6] se observa que, si bien ninguna correlación es superior a 0.5, existen ciertos pares de datos que definitivamente no se corre-

lacionan y otros que tienen cierta cercanía.

Tras lo anterior, surge naturalmente la necesidad de observar las correlaciones temporales de un mismo parámetro. Con eso en mente, se despliegan en la figura [3] las autocorrelaciones de los parámetros considerando desplazamientos temporales de hasta 10 días. Ahí se aprecia de modo más evidente que los parámetros tienen alta autocorrelación y que por lo tanto la temporalidad no puede ser ignorada, siendo propicio para esto el modelo escogido.

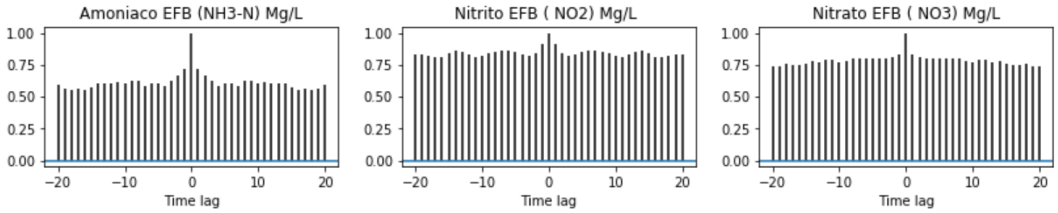

Figura 3: Autocorrelación para las mediciones de amoniaco, nitrito y nitrato

Por otro lado, una opción interesante es desagregar la información contenida en la correlaciones entre 2 parámetros. Esto se puede lograr controlando por una tercera variable, fijando un rango por ejemplo. Existía la hipótesis de que a ciertas temperaturas, los valores de nitrito y nitrato se relacionaban fuertemente. Esto se pudo comprobar al ver que en el rango de los 23º a 24º, la correlación es de 0.8, muy por sobre el 0.5 reportado originalmente sin considerar la temperatura. Esto da luces de que efectivamente existe una mayor correlación entre la concentración de ambos compuestos, pero que es difícil de apreciar si se consideran rangos amplios de temperatura. Resultados similares ocurren con el pH, donde ciertos intervalos aumentan la correlación.

3.3. Descripción de los datos

Por último, una descripción de los datos desde el punto de vista de las medidas estadísticas comunes, entrega que las medias, mínimos y máximos de los datos medidos en las piscinas, están acordes a los rangos esperados. Sin embargo, muestran que existe en ciertas mediciones una desviación muy alta, como es el caso del nitrato, donde para una media de 17.6 mg/L en una de las piscinas, se obtuvo una desviación estándar de 12.9. Lo anterior puede anticipar la gran variación que puede tener nuestro modelo a la hora de predecir, lo que sugiere la necesidad de más y mejores mediciones para poder predecir con certeza los valores esperados.

3.4. ARMA

El modelo ARMA consta de tres elementos que permiten predecir una variable: autorregresión, ventana móvil y error. Cada uno aporta de una forma particular a obtener resultados que pueden ser muy cercanos a la realidad en muchos casos.

La autorregresión se refiere al efecto inercial de la propia variable, indica en qué medida tiende a mantenerse estable, cuánto afectan sus valores anteriores en el presente.

La ventana móvil, asume que el efecto que realizan las entradas en el valor actual de y se realiza de dos formas: influyendo sobre valores pasados de la propia variable y, y también de forma directa sobre su valor actual. Este efecto, se considera por un cierto período en el tiempo, a esto se le llama ventana móvil. En este caso se entiende que las entradas son algunos de las otras variables medidas.

Por último, se incluye un término de error para asumir que pueden existir perturbaciones no consideradas en el modelo, que lo hacen diferir de la realidad. Aquí se agrupa el efecto de las variables no introducidas y también de otras cuyos datos no se consideran, como por ejemplo, la cantidad de peces o el nivel de recambio del agua.

Los usos que se le pueden dar a modelos tipo ARMA son muy variados, áreas de predicción y control los ocupan de forma continua. En este último ámbito, se suelen usar las entradas para modelar aquello que se puede manejar, y agrupar los términos restantes en el error o en otros términos separados. Esto tiene sentido porque lo que se intenta es definir ese vector de control que permitirá manejar la salida. En casos de predicción, la separación se hace de forma distinta puesto que no hay entradas. Aquí, cobra sentido la idea de separar las variables que afectan de forma directa y en mayor medida en la salida, para considerar que las restantes no son más que ruido o un error pequeño, no modelado en detalle.

El modelo ARMA se implementó con los siguientes valores para los parámetros mencionados anteriormente: $p = 2$, $q = 3$, $M = 2$, ϵ: $\mathcal{N}(0, 1.3)$. Los parámetros recién mencionados y las tres mediciones que se usaron como entrada, se obtuvieron iterando con distintos números y conjuntos de variables para ver cuál tenía mejores resultados.

Para obtener la *likelihood* de los parámetros, se asumió independencia y se obtuvo la probabilidad del dato real de nitrito, dada la estimación realizada a partir de cada conjunto de datos.

3.5. Metrópolis-Hastings

Luego, se ocupó el método de muestreo Metrópolis-Hastings para la obtención de *sets* de coeficientes para ARMA. Esto se realizó utilizando como *jumpling distribution* una gaussiana centrada en la muestra anterior (una para cada coeficiente) y con una desviación estándar alta para lograr que recorra bien el espacio de posibles coeficientes. Como p se usó directamente la log-likelihood de ARMA, ya comentada. El *alpha* usado para aceptar la muestra fue el siguiente:

$$\alpha = min(exp(p^*(x^*) + q(x_t \mid x^*) - p^*(x_t) + q(x^* \mid x_t)), 1)$$

4. Resultados y análisis

Se muestran algunos gráficos que comparan la predicción realizada con los valores reales. Hay diferencias muy significativas en la decisión de qué entradas usar o qué parámetros elegir para el modelo ARMA.

Fue posible hallar dos combinaciones de parámetros y entradas que logra predecir de manera bastante certera el valor de los nitritos. Es cierto que el ajuste no es perfecto ni modela cada subida o bajada, pero si muestra un buen seguimiento de tendencias que es de gran importancia. De hecho, un seguimiento más cercano podría ser muestra de un sobre ajuste.

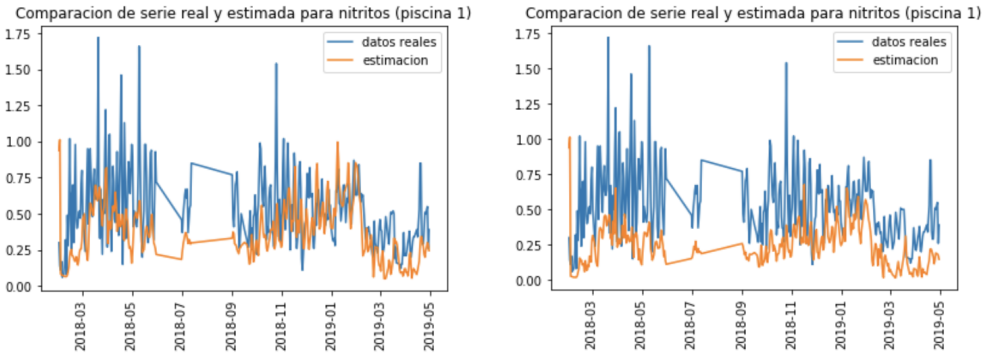

Figura 4: Serie estimada usando nitrato y temperatura (izquierda), nitrato y pH (derecha)

Las mejores predicciones se obtuvieron usando como entrada dos datos anteriores de nitritos, y tres datos anteriores de temperatura y nitrato.

La inclusión de más información en tiempo o más variables provoca una complejización del problema que afecta en la calidad de los resultados de la predicción. En los anexos [7] se muestra un ejemplo de ello, en que se usaron 5 períodos y se consideran 5 entradas.

Sería prudente realizar un estudio más profundo de la química que opera en las piscinas, para entender cómo es la relación intertemporal entre parámetros. Esto permitiría darle adecuados parámetros iniciales al modelo para que luego itere perfeccionándolos. Además, daría luces de formas de combinación de parámetros no consideradas aquí o sobre la necesidad de incluir algún grado de no linealidad.

Por otro lado, la hipótesis de que la correlación entre los parámetros es relevante, pero lo es aún más si se consideran correlaciones con otros parámetros fijos, se comprueba. Se realizó el mismo experimento, con otro parámetro altamente correlacionado al nitrito, el amonio, pero que no presenta una correlación mayor a temperaturas controladas. Los resultados para ese parámetro fueron muy lejanos a los buenos resultados obtenidos anteriormente, lo que comprueba la relevancia de esa información, como se muestra en el anexo [8].

5. Conclusiones

En el presente proyecto se planteó la posibilidad de predecir el valor de los nitritos en piscinas acuícolas a partir de datos pasados y de otras variables más fáciles de medir. Para ello fue necesario hacer un análisis de los datos que permitiera entender a grandes rasgos las principales características de comportamiento y relaciones existentes entre las variables. Luego, fue posible plantear un modelo que tomara en cuenta los hallazgos y luego definir los detalles del mismo.

El estudio de los datos sirvió para entender qué experimentos tiene sentido hacer y cuáles no, qué metodologías pueden aportar información relevante y cuales no son adecuadas para este tipo de datos. Se optó por las metodologías que consideran la temporalidad de los datos *versus* las que los toman como mediciones independientes.

Por otro lado, se observó que un análisis exploratorio inicial aporta perspectivas sobre la estructura de los datos, pero que existen relaciones intrínsecas de los mismos que pueden aparecer al estudiarlos más detenidamente, como es la correlación entre variables fijando otras. También que un buen análisis permite escoger parámetros adecuados que tienen una alta incidencia en los resultados.

Se pudo ejecutar una rutina de Metrópolis-Hastings que opera buscando parámetros para el modelo ARMA propuesto. Se obtuvieron resultados en un principio lejanos a la realidad pero que luego con un análisis detenido de los datos se ajustaron mucho más, llegando a obtener series de tiempo bastante ajustadas a la real, que pueden servir de referencia a la hora de predecir datos futuros.

El método propuesto resultó ser sumamente sensible a las correlaciones entre los datos y a la temporalidad, como se había teorizado en un principio. De este modo, la comprensión de los datos fue fundamental para escoger el modelo, seleccionar los parámetros, y finalmente decidir qué experimentos realizar.

Inclusión de estudios químicos sobre las relaciones entre varias, o la temporalidad de los efectos entre una y otra podrían aportar de forma significativa a idear modelos que capten mejor la realidad. De igual forma, un aumento de los datos podría significar un crecimiento en robustez de los resultados que pueden verse como pasos a seguir con miras a alcanzar resultados útiles para la empresa.

Por último señalar que con pocos datos, técnicas relativamente sencillas y algoritmos sumamente estudiados, se logró aplicar lo aprendido a un problema real como es predecir el nivel de nitritos en una planta de crianza de esturiones, lo que puede significar grandes ahorros de costos para el proyecto en cuestión.

Referencias

[1] Ling, Zhu, and Yee. "Diagnostic Checking for Non-stationary ARMA Models with an Application to Financial Data."North American Journal of Economics and Finance 26.C (2013): 624-39. Web.

[2] Mcleod, and Zhang. "Faster ARMA Maximum Likelihood Estimation.Çomputational Statistics and Data Analysis 52.4 (2008): 2166-176. Web.

[3] Ravishanker, Nalini, and Bonnie Ray. "Bayesian Analysis of Vector ARMA Models Using Gibbs Sampling."Journal of Forecasting 16.3 (1997): 177-94. Web.

[4] Dong and Peng. "Principled missing data methods for researchers". SpringerPlus 2.1 (2013): 222. Web.

Anexos

Se incluyen 2 cuadernillos (*jupyter notebooks*) a modo de anexo con el análisis de los datos, y el modelo junto con su implementación.

El análisis de datos presente en `EDA.ipynb` incluye:

- Histogramas para todos los parámetros de la piscina 1.
- Histogramas con ajuste normal.
- Series de tiempo de todos los parámetros.
- Series de tiempo promediadas por mes.
- Matriz de correlaciones.
- Gráficos de parámetro *vs* parámetro que muestran correlación.
- Gráficos de autocorrelación para todos los parámetros.
- Descripción de los datos.
- Análisis de correlación de 3 parámetros
- Exploración respecto al mecanismo de *missing data*.

Parte del desarollo del modelo se incluye en `modelo1.pdf` y `modelo2.pdf` que incluyen distintas iteraciones del código utilizado para ARMA, Metropolis-Hastings, algunas visualizaciones, entre otros.

Parámetro	**Unidad**	**Rango Equipo**	**Rango Cultivo**	**Óptimo**
Temperatura	°Celsius	0 a 30	8 a 25	20
Amonio (NH_4)	mg/L	0 a 300	<3	<3
Amoniaco (NH_3)	mg/L	0 a 300	<1	<1
Nitrito (NO_2)	mg/L	0 a 1.15	<3	<3
Nitrato (NO_3)	mg/L	0 a 100	70	70
pH	Unidad	0 a 14	7 a 8	7

Tabla 1: Parámetros del agua con unidades y rangos

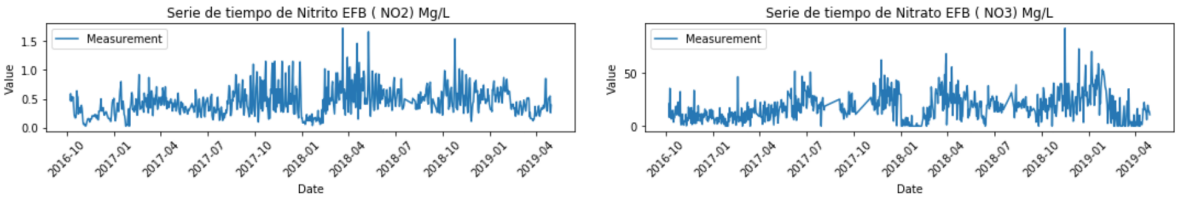

Figura 5: Series de tiempo diarias para las mediciones de nitrito y nitrato

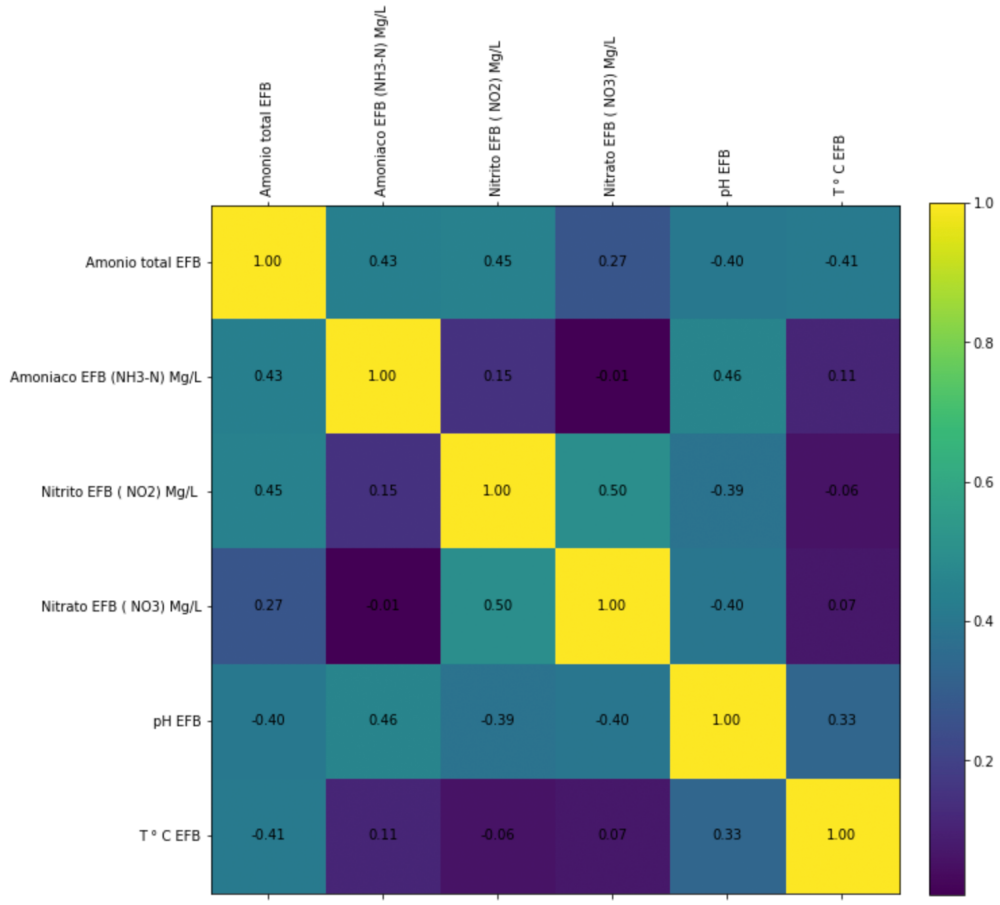

Figura 6: Matriz de correlaciones entre parámetros

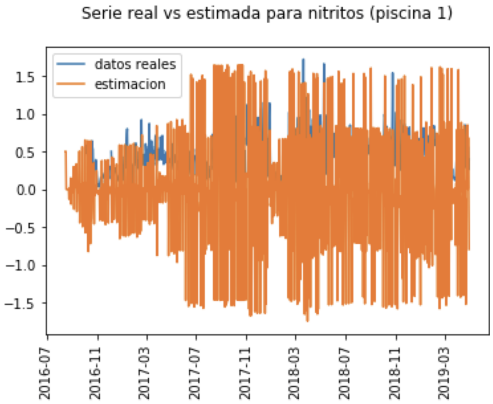

Figura 7: Resultados para la serie de tiempo predicha con exceso de parámetros.

Figura 8: Resultados para la serie de tiempo con amonio y temperatura.