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

Entrega 1 Proyecto Semestral

11 de junio 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 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 a razón 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. Como los datos fueron obtenidos de forma manual, la posibilidad de error es considerable y además, faltan datos.

En la tabla [1] se muestran los principales parámetros medidos actualmente con sus rangos y óptimos.

Parámetro	**Unidad**	**Rango Equipo**	**Rango Cultivo**	**Óptimo**
Temperatura	ºCelsius	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

Con esos datos a disposición, uno de los posibles resultados más valiosos, es la estimación de la concentración de nitrito en el agua, puesto que es el sensor más costoso de obtener.

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 autoregresivo, particularmente modelos ARMA (*Autoregressive Moving Average Models*).

2. Marco Teórico

Luego de un análisis acabado de los datos fue posible constatar que el uso de los datos 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 mismo 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:

$$A(q)y[k] = B(q)u[k] + C(q)e[k]$$

En este modelo, se reciben entradas u que, junto a los valores anteriores de y y el error e definen cuál va a ser el nuevo valor para y. k representa la temporalidad, que en general es

discreta. A, B y C son los coeficientes que permiten ajustar el modelo.

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, considerando como entrada algunos de los demás parámetros y como ruido los demás. Se asume que el ruido es *gaussiano*.

Se busca encontrar los coeficientes de A, B, y C. La forma en este caso, será la maximización de la *likelihood*. Como el espacio de coeficientes, no es posible usar directamente algún método de optimización, razón por la cual el uso de muestreos inteligentes aparece como buena alternativa. De hecho, la literatura muestra que métodos como Gibbs-Sampling o su versión general, Metrópolis-Hastings pueden ser ocupados para esto [3]. En esta primera versión se usó Metrópolis-Hastings para obtener los coeficientes, pero se mantiene abierta la posibilidad de probar con Gibbs-Sampling para comparar resultados.

La posibilidad de completar datos faltantes en la base de datos sigue vigente porque las filas que presentan la totalidad de sus datos son en realidad menos de un tercio del total de datos. Así, los métodos gráficos u otros surgen como alternativa atractiva para aumentar la información disponible para encontrar los coeficientes.

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.

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, puesto que son los más correlacionados, el de nitrito y el de nitrato, junto con el ajuste normal. Para poder mostrar la distribución y el histograma, se presenta el histograma de densidades.

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

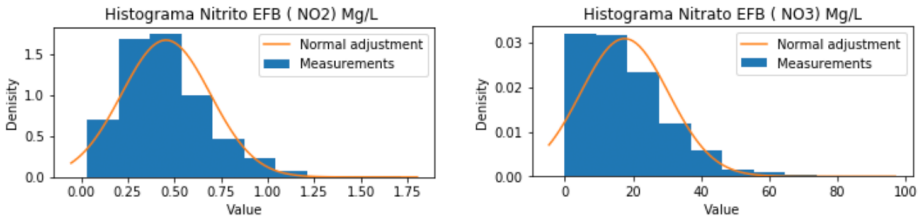

Figura 1: Histograma para mediciones de nitrito y nitrato

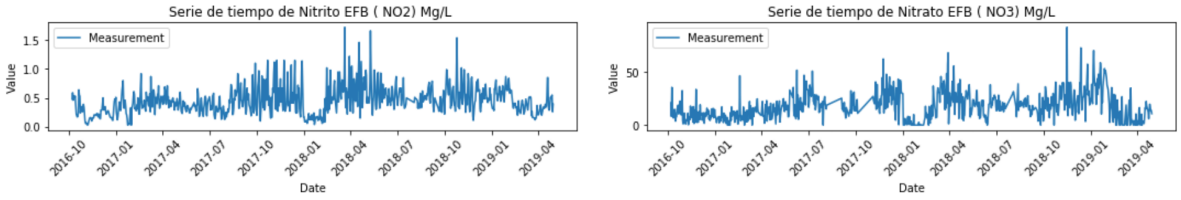

Figura 2: Series de tiempo diarias para las mediciones de nitrito y nitrato

adelante no es excesivamente alta.

En segundo lugar, se visualizaron los parámetros como series de tiempo, lo que es de especial interés dado el modelo a utilizar. Dos de las más notables se muestran en la imagen [2], las cuales nuevamente no entregan demasiada información respecto a una potencial estructura.

Se optó por mostrar series de acuerdo a promedios de las mediciones en un mes [3] 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.

3.2. Correlaciones

En el análisis de la correlación se encontraron parámetros relacionados entre sí, los cuales confirmaron algunas hipótesis sobre el comportamiento de ciertos compuestos químicos,

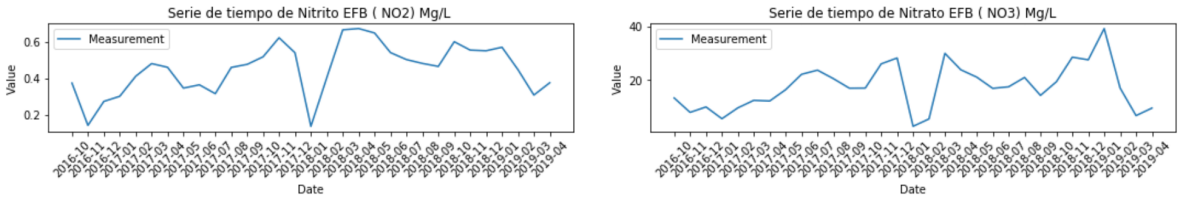

Figura 3: Series de tiempo promediadas por mes para las mediciones de nitrito y nitrato

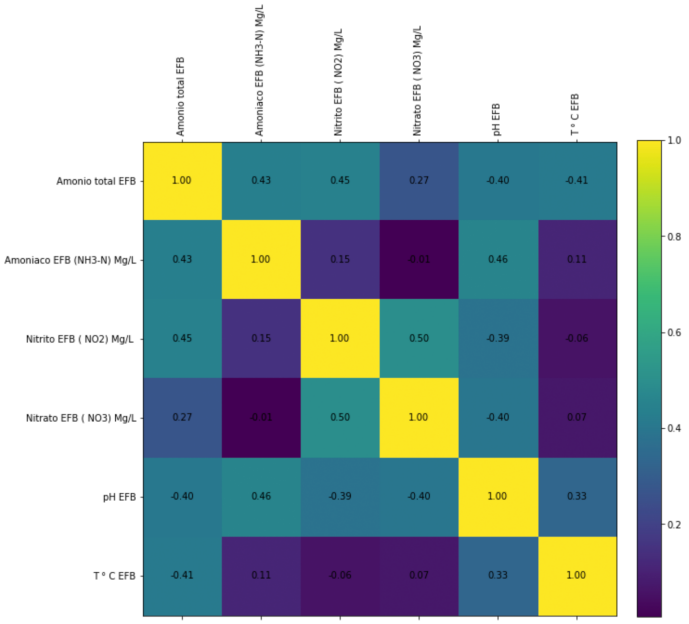

Figura 4: Matriz de correlaciones entre parámetros

como por ejemplo que por su composición, el nitrito y nitrato deben estar correlacionados.

En la matriz de correlaciones mostrada en [4] se observa que si bien ninguna correlación es superior a 0.5, si existen ciertos pares de datos que definitivamente no se correlacionan 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 [5] 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.

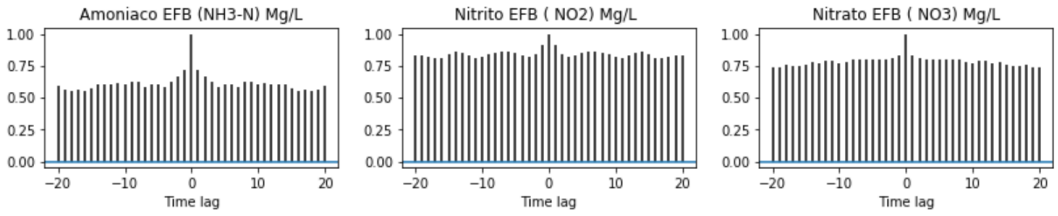

Figura 5: Autocorrelación para las mediciones de amoniaco, nitrito y nitrato

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 se implementó siguiendo la siguiente fórmula:

$$y[k] = a_2 * y[k-2] + a_1 * y[k-1] + b_2^n * n[k-2] + b_1^n * n[k-1] + b_0^n * n[k]$$

$$+b_2^p * p[k-2] + b_1^p * p[k-1] + b_0^p * [k] + b_2^m * m[k-2] + b_1^m * m[k-1] + b_0^m * m[k] + e[k]$$

Donde y =nitrito, n =nitrato, p =pH, m =amonio. El error se consideró gaussiano y afecta solo al instante k directamente.

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 set 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:

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

4. Resultados y análisis

Los resultados obtenidos hasta el momento son muy básicos y todavía no se acercan a lo esperado pero ya muestran un funcionamiento de la operatoria de cálculo detrás del modelamiento. Se muestran estimaciones para las dos piscinas de los nitritos, junto con los datos

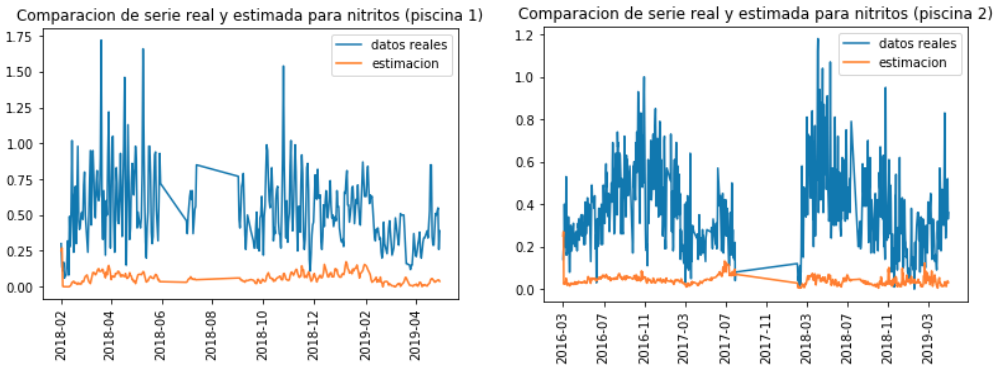

Figura 6: Resultados para la serie de tiempo predicha en cada piscina

reales.

Estas predicciones se obtuvieron usando como entrada, dos datos anteriores de nitritos, y tres datos anteriores de pH, amonio y nitrato. Se espera que al agregar las demás variables y realizar iteraciones *Metrópolis-Hastings* hasta lograr convergencia se obtengan series más cercanas a la realidad.

Los resultados muestran una mejora significativa de la likelihood desde el punto inicial en -1247 hasta el final en -950, pero en todo caso muy lejano de lo que se espera llegar.

5. Conclusiones

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. Se comprendió que la correlación entre datos de entrada y salida del filtro biológico (las primeras 5 columnas y las siguientes 5) es muy alta, razón que no permite asumir su independencia para aumentar los datos sin más trabajo.

Se pudo ejecutar una rutina de Metrópolis-Hastings que opera buscando parámetros para el modelo ARMA propuesto. Se obtuvieron resultados lejanos a la realidad pero que muestran un código operativo al que le falta ajustarse y operar por más tiempo para acercarse a la realidad.

El mayor avance es en realidad la comprensión de una metodología conjunta que permite lograr predicciones y la programación de una versión básica de esta. Con eso, ya es posible ir perfeccionando la técnica, agregando más datos y robusteciendo el método para lograr

resultados útiles.

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.

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.