



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

## Propuesta Proyecto Semestral

14 de mayo de 2019

Raimundo Herrera - Manuel Vial

---

### 1. Introducción

Existe un proyecto en ejecución de crianza de esturiones en Parral, región del Maule. Estos viven en piscinas con un sistema llamado *de recirculación*, que opera reutilizando el agua luego de una limpieza de la misma con filtros.

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 al menos 5 parámetros, 6 veces al día de forma continuada. Como éstos fueron tomados de forma manual, es altamente probable que tengan errores de distintos tipos.

Con el paso de los años se ha hecho notoria la necesidad de incorporar nuevas mediciones. En la tabla [??] se muestran los principales parámetros medidos actualmente. De acuerdo a lo significativo que sea cada parámetro se discriminará si utilizar únicamente aquellos de los que se tenga registro desde el primer año o aquellas mediciones más recientes pero con menos datos.

Parámetro	Unidad	Rango Equipo	Rango Cultivo	Óptimo
Temperatura	°Celsius	0 a 30	8 a 25	20
Oxígeno	mg/L	0 a 60	5 a 9	9
Saturación	%	0 a 600	90	90
Amonio ( $NH_4$ )	mg/L	0 a 300	<3	<3
Nitrito ( $NO_2$ )	mg/L	0 a 1.15	<3	<3
Nitrato ( $NO_3$ )	mg/L	0 a 100	70	70
Alcalinidad ( $CaCO_3$ )	mg/L	0 a 500	>150	>150
Dureza ( $CaCO_3$ )	mg/L	0 a 250	>70	>70
pH	Unidad	0 a 14	7 a 8	7
Potencial ORP	mV (mini Volt)	0 a 300	250	250
Presión de gases totales	%	0 a 600	100	100
Dióxido de Carbono ( $CO_2$ )	mg/L	0 a 50	2	2
DBO	mg/L	0 a 500	<35	4

Tabla 1: Parámetros del agua con unidades y rangos

Los datos están disponibles y el permiso para su uso ya fue conseguido. Se espera en el transcurso de las siguientes semanas estudiar su forma y orden.

## 2. Marco Teórico

Para el análisis de relaciones entre variables, se propone como método base el uso de regresión lineal bayesiana. Éste sería un punto de partida y serviría como comparación frente a métodos más sofisticados que pudieran entregar resultados más precisos en caso de éxito.

Una de las particularidades de la regresión bayesiana es que puede utilizar alguno de los diversos métodos de *sampling* que hemos revisado durante el curso para estimar la distribución de la posterior de los parámetros que se deseen obtener. En este caso en particular, la intención es utilizar métodos como *Gibbs Sampling* [?] para alimentar la regresión. Lo anterior dado que procedimientos como éste permiten generar muestras aún cuando el dominio del problema sea de alta dimensionalidad, por medio de condicionamientos en otros parámetros, algo muy similar a lo que se pretende, dada la cantidad de parámetros con los que se cuenta.

Por otra parte, una de las variantes a explorar del problema en cuestión, toma en cuenta estimar el comportamiento como series de tiempo de los parámetros medidos. Lo anterior con el objetivo de mitigar riesgos, prevenir muertes masivas de peces, etc. Se propone que para dicha predicción se utilicen procesos de regresión gaussianos [?], [?]. En línea con lo investigado hasta el momento, la utilización de métodos de Monte Carlo en cadenas de Markov, como el de Metropolis-Hastings aparecen como candidatos naturales para encontrar los

parámetros de dichas series de tiempo.

Dadas las características de los datos a nuestra disposición, una posibilidad a explorar es la de completar los datos faltantes por medio de alguno de los métodos estudiados. Como se explicó anteriormente, existe información desde el comienzo de las mediciones, sin embargo, con el tiempo se han agregado nuevos parámetros al conjunto de datos por lo que puede ser necesario estudiar las distribuciones y comportamiento de las mediciones por sí solas, para determinar si es posible ajustar distribuciones y generar datos para las muestras que no tengan ciertas mediciones.

### 3. Conclusiones

Hay dos posibles resultados que podrían ser de gran utilidad:

- encontrar alguna relación entre los parámetros existentes, que permita no medir alguno de ellos.
- encontrar ciclos o indicadores de que un parámetro relevante va a cambiar, para tomar por anticipado las medidas necesarias para evitar dicho cambio o paliar sus consecuencias.

Uno de los beneficios directos del primer resultado, como se comentó anteriormente, es que podría reducir significativamente los costos del proyecto de automatización en cuanto existen ciertas mediciones para las cuales se requiere incurrir en grandes gastos. Por ejemplo, si se lograra encontrar una forma robusta de obtener el valor de la concentración de nitritos en las piscinas usando los demás parámetros, se podría ahorrar mucho, dado que los sensores para dicha medición son muy costosos.

En cuanto al segundo resultado, éste podría ayudar a disminuir la cantidad de muertes de peces por imprevistos. Actualmente si ciertos parámetros abandonan los rangos deseados, la cantidad de peces que mueren es del orden de miles.

El análisis anterior muestra potenciales beneficios para la crianza de esturiones. No obstante, es importante señalar que este trabajo serviría también para el análisis comparativo de los modelos propuestos en cuanto a su utilidad predictiva en este marco en particular, y la eventual generalización de los mismos para aquellos problemas donde los datos tengan ciertas propiedades en común con éste.

Una de las principales dificultades se relaciona con la limpieza de los datos obtenidos. Durante los 5 años de toma de datos, se hizo manualmente. Se espera que ciertas mediciones estén incompletas, con errores, o con valores poco fidedignos. El desafío es lograr obtener resultados significativos a pesar de esas dificultades.

## Referencias

- [1] A. Smith, G. Roberts. *Bayesian computation via the Gibbs sampler and related Markov chain Monte Carlo methods*. J. Roy. Statist. Soc. Ser. B 55, 3-23.
- [2] R. M. Neal. *Monte Carlo implementation of Gaussian process models for Bayesian regression and classification*. Technical Report 9702, Dept. of Statistics, Univ. of Toronto, 1997. Available at <https://arxiv.org/pdf/physics/9701026>
- [3] C. Williams, D. Barber. *Bayesian classification with Gaussian processes*. IEEE Transactions on Pattern Analysis and Machine Intelligence, 1998, vol. 20, no 12, p. 1342-1351.