UNIVERSIDAD DE PIURA

FACULTAD DE INGENIERÍA

Análisis e identificación de las variables de mayor influencia para el crecimiento de los langostinos

Tesis para optar el Título de
Ingeniero Industrial y de Sistemas

David Percy Sojo Chero

Asesor(es):

Dra. Ing. Ana Valeria Quevedo Candela; Dra. Ing. Bertha Susana Vegas Chiyón

Piura, setiembre de 2020

UNIVERSITAS STUDIORUM PIURENSIS

Agradecimientos

Agradezco a Dios por permitirme tener y disfrutar de mi familia, gracias a mi familia quienes me han apoyado de manera incondicional en cada decisión, momento difícil y proyecto, por permitirme cumplir satisfactoriamente el desarrollo del presente trabajo. Gracias a la familia García Calopiña por acogerme, de la manera más amable en su hogar, durante muchas horas lo que permitió el desarrollo de buena parte de tesis.

Agradezco a mi asesora, Dr. Ing. Valeria Quevedo, quien ha sido una excelente instructora y mentora, sin duda un actor fundamental para la conclusión exitosa del presente trabajo. Gracias a mi co-asesora, Dr. Ing. Susana Vega, quien me permitió formar parte de este proyecto y poder aportar con mi trabajo al mismo.

Gracias al programa nacional de becas y crédito educativo (PRONABEC) por permitirme acceder, de forma gratuita, a una educación superior en una casa de estudio tan buena como lo es la Universidad de Piura.

El presente estudio tomó buen tiempo y requirió de muchas horas de trabajo que fueron acompañadas del apoyo y ánimo de todas las personas de mi entorno; amigos, familiares y maestros a quienes estoy profundamente agradecido y que sin ellos esto no sería posible.

Este trabajo ha sido financiado por el Proyecto Concytec – Banco Mundial "Mejoramiento y Ampliación de los Servicios del Sistema Nacional de Ciencia Tecnología e Innovación Tecnológica" 8682-PE, a través de su unidad ejecutora Fondecyt, bajo el contrato E041-2018-01-BM.

CONCYTEC | FONDECYT Becas y Financiamiento del Concytec

BANCO MUNDIAL
BIRF • AIF | GRUPO BANCO MUNDIAL

UNIVERSITAS STUDIORUM PIURENSIS

Resumen Analítico-Informativo

Análisis e identificación de las variables de mayor influencia para el crecimiento de los langostinos

David Percy Sojo Chero.
Asesor(es): Dra. Ing. Ana Valeria Quevedo Candela.
Dra. Ing. Bertha Susana Vegas Chiyón.

Tesis.
Ingeniero Industrial y de Sistemas.
Universidad de Piura. Facultad de Ingeniería.
Piura, Setiembre de 2020

Palabras claves: Crecimiento de los langostinos, limpieza de datos, función Gompertz (Von Bertalanffy), regresión no lineal, regresión lineal múltiple, validación cruzada, y análisis de sensibilidad.

Introducción: El incremento del consumo de recursos hidrobiológicos (pesquero, acuícola, etc.) genera el incremento de las actividades de pesca y acuicultura, produciendo un crecimiento en los mercados proveedores para abastecer esta demanda. La actividad de la acuicultura, en respuesta al entorno de crecimiento, tiene la necesidad de mejorar y ser más eficiente, ya que actualmente en el Perú se realiza de manera empírica. El presente estudio se refiere a la mejora de la crianza del langostino en piscinas de agua dulce. Para ello, mediante la aplicación de herramientas estadísticas, se propone identificar los factores de mayor influencia en el crecimiento de los langostinos durante su etapa de engorde, para focalizar los esfuerzos de mejora y control en estos factores y obtener mejores resultados.

Metodología: Este estudio propone identificar dichos factores con ayuda de un modelo de dos fases. La primera fase pretende modelar el crecimiento medio del langostino de cada piscina a través de un modelo no-lineal. Mientras que la segunda fase pretende identificar aquellos factores que, expliquen la variación respecto al crecimiento medio esperado a través de un modelo de regresión múltiple. Sin embargo, para iniciar con dicho análisis es necesario realizar un tratamiento de limpieza de los datos para evitar trabajar con datos errados.

Resultados: Tomando como partida las variables que realmente se miden y se registran, se pueden identificar aquellas de mayor influencia en el crecimiento del langostino. A partir del modelado de datos de este estudio se identificó que la temperatura, el oxígeno, pH, entre otras variables.

Conclusiones: Se concluyó que, las variables identificadas como más importantes, tienen un notable efecto con respecto al crecimiento medio, es decir, dependiendo de la variación generada en estas variables se puede conseguir un efecto positivo o negativo sobre la curva media de crecimiento del langostino.

Fecha de elaboración del resumen: 28 de Agosto de 2020

Analytical-Informative Summary

Análisis e identificación de las variables de mayor influencia para el crecimiento de los langostinos

David Percy Sojo Chero.
Asesor(es): Dra. Ing. Ana Valeria Quevedo Candela.
Dra. Ing. Bertha Susana Vegas Chiyón.

Tesis.
Ingeniero Industrial y de Sistemas.
Universidad de Piura. Facultad de Ingeniería.
Piura, Setiembre de 2020

Keywords: Shrimp growth, data cleansing, Gompertz function (Von Bertalanffy), nonlinear regression, multiple linear regression, cross-validation, and sensitivity analysis.

Introduction: The increase in the consumption of hydrobiological resources (fishing, aquaculture, etc.) generates an increase in fishing and aquaculture activities, producing growth in the supplier markets to supply this demand. Aquaculture activity, in response to the growth environment, needs to improve and be more efficient, since currently in Peru it is carried out empirically. The present study refers to the improvement of shrimp farming in freshwater pools. To do this, through the application of statistical tools, it is proposed to identify the factors with the greatest influence on the growth of prawns during their fattening stage, in order to focus improvement and control efforts on these factors and obtain better results.

Methodology: This study proposes to identify these factors with the help of a two-phase model. The first phase aims to model the average growth of the shrimp in each pool through a non-linear model. While the second phase aims to identify those factors that explain the variation with respect to the expected average growth through a multiple regression model. However, to start with such analysis it is necessary to carry out a data cleaning treatment to avoid working with erroneous data.

Results: The starting point are the variables that are actually measured and recorded, those with the greatest influence on shrimp growth can be identified. From the data modeling of this study, it was identified that temperature, oxygen, pH, among other variables.

Conclusions: It was concluded that the variables identified as the most important, have a notable effect with respect to the average growth, that's mean, depending of the variation generated in these variables, we can achieve a positive or negative effect on the average growth curve of the shrimp.

Summary date: August 28, 2020.

Tabla de contenido

Lista de figuras

Lista de Tablas

UNIVERSITAS STUDIORUM PIURENSIS

Introducción

El incremento del consumo de recursos hidrobiológicos (pesquero, acuícola, etc.) genera el incremento de las actividades de pesca y acuicultura, produciendo un crecimiento en los mercados proveedores para abastecer esta demanda. La actividad de la acuicultura, en respuesta al entorno de crecimiento, tiene la necesidad de mejorar y ser más eficiente.

El presente estudio se refiere a esta actividad y en específico a la mejora de la crianza del langostino en piscinas de agua dulce. Para ello, mediante la aplicación de herramientas estadísticas, este estudio propone identificar los factores de mayor influencia en el crecimiento de los langostinos durante su etapa de engorde, para focalizar los esfuerzos de mejora y control en estos factores y obtener mejores resultados.

Para el desarrollo de este trabajo se recurrió a una data facilitada por la empresa ECOSAC, agroexportadora ubicada a tan solo 8 km de la ciudad de Piura que cuenta con una línea de productos integrada por uva de mesa, palta, mango, pimiento, langostinos, etc. La data es un registro del peso de los langostinos, características de las siembras y parámetros como temperatura, oxígeno, pH, alcalinidad entre otros, que se encuentran ordenados por sus respectivas piscinas y fechas.

Este estudio propone identificar dichos factores con ayuda de un modelo de dos fases. La primera fase pretende modelar el crecimiento medio del langostino de cada piscina a través de un modelo no-lineal. Mientras que la segunda fase pretende identificar aquellos factores que expliquen la variación respecto al crecimiento medio esperado a través de un modelo de regresión múltiple. Sin embargo, para iniciar con dicho análisis es necesario realizar un tratamiento de limpieza de los datos para evitar trabajar con datos errados.

UNIVERSITAS STUDIORUM PIURENSIS

Capítulo 1

Antecedentes

En este capítulo describirá el entorno mundial y nacional del sector pesquero y acuícola para contextualizar el entorno en el que se desarrolla el presente estudio. Se iniciará con una visión general de las actividades y sobre la tendencia mundial, para luego pasar a describir la acuicultura de langostinos en el Perú.

1.1. Situación actual de la pesca y la acuicultura

Durante los últimos años el sector pesquero y acuícola han crecido de manera significativa, para el año 2018 alcanzó un crecimiento del 39.73% (INEI, 2019), aunque la especie que más contribuye a este crecimiento es la anchoveta para consumo indirecto, se tiene por otro lado que cierta parte de este crecimiento es explicado por el aumento de captura de especies para consumo directo, que para el 2018 fue de 6.42% (INEI, 2019), entre estas especies tenemos: conchas de abanico, langostino, anchoveta, atún, pota, perico y liza.

Según Ministerio de la producción (2017) y al margen del reglamento Ley N° 27460-Ley de Promoción y Desarrollo de la acuicultura, determinan a esta como el conjunto de actividades tecnológicas orientadas al cultivo y crianza de especies acuáticas de forma total o parcial en ambientes hídricos naturales o artificiales seleccionados y controlados, como aguas marinas, dulces o salobres.

Según se muestra en la Figura 1 en el futuro el crecimiento pesquero derivará principalmente de la actividad de la acuicultura, debido al aumento de la práctica de esta actividad y la disminución de la pesca por captura a causa del aumento de las restricciones, a causa de la preocupación por la depredación de las especies marinas. Este ascenso de la producción acuícola se puede notar en las tasas de crecimiento de consumo *per capita* por especies. En el 2015, el consumo mundial *per capita* llego a los 7.8 kg lo cual representa el 38% del total, como se muestra en la Figura 2, la cantidad de pescado para el consumo humano es proveído por la pesca y la acuicultura, esta última se encuentra en un ascenso progresivo que en los últimos 20 años ha

tomado una mayor relevancia en cuanto al aprovisionamiento de pescado para el consumo directo por humanos. (FAO, 2019)

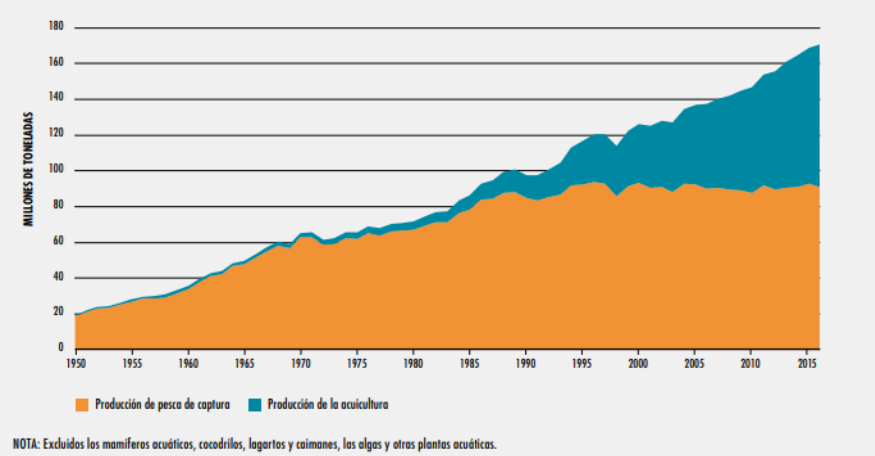


Figura 1. Producción mundial de la pesca de captura y agricultura 2015
Fuente: Estado mundial de la pesca y la acuicultura (2018)

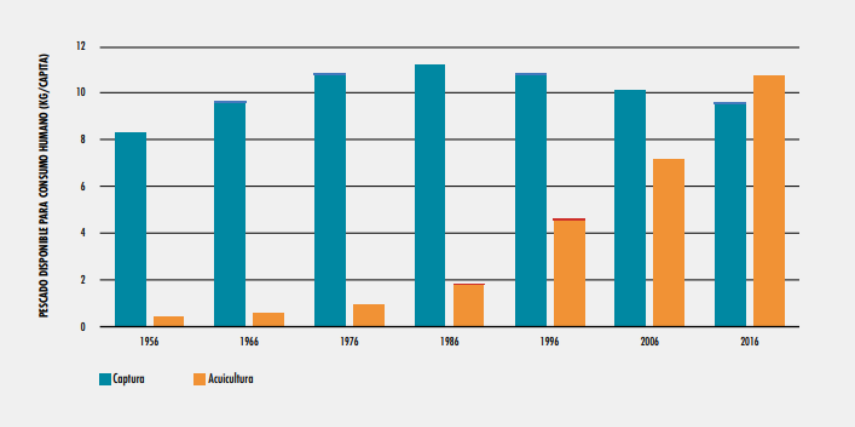


Figura 2. Contribución relativa de la acuicultura y la pesca por captura para el consumo humano
Fuente: Estado mundial de la pesca y la acuicultura (2018)

El crecimiento de la actividad de la acuicultura se evidencia en la gran variedad de especies acuáticas, cerca de 580, que se cultivan en el mundo y que para el 2016 se alcanzó un volumen récord de 80 millones de toneladas a nivel mundial, lo que representaría casi un 50% del pescado destinado al consumo directo a nivel mundial. (FAO, Organización de las Naciones Unidas para la Alimentación y la Agricultura, 2019)

En la actualidad, la acuicultura produce cerca de 73.8 millones TM, lo cual representa la mitad del consumo directo humano. El continente asiático es el mayor productor y, en específico,

China es el país con mayor producción (45.4 millones de TM), responsable de más del 60% de la producción mundial. (Departamento de inteligencia de Mercados, 2017)

Para reconocer la participación de la producción acuícola del Perú en el mundo se pueden recurrir a datos de años anteriores debido a que los datos son relativamente similares. Según se muestra en la Figura 3, en el 2016, la acuicultura peruana produjo lo 100 000 TM en total de todas sus variedades cultivadas, lo que representó el 0.16% de la producción de China, el mayor productor acuícola a nivel mundial, y el 10% de la producción acuícola en Chile, el mayor productor acuícola de Sudamérica. Mientras que nuestra participación a nivel mundial fue en ese año del 0.1 %. (ComexPerú, 2018)

	1980	1990	2000	2005	2016	Participación en la producción mundial (2016)	Participación en la producción de SA (2016)
Mundo	**7,189**	**16,850**	**41,724**	**57,820**	**110,208**	**100.0%**	**-**
China	2,660	7,963	28,460	37,615	63,722	57.8%	-
Suramérica (SA)	17	186	744	1,247	2,321	2.1%	100%
Chile	2	70	425	739	1,050	1.0%	45%
Brasil	4	20	172	258	581	0.5%	25%
Ecuador	10	78	61	139	451	0.4%	19%
Perú	1	5	7	26	100	0.1%	4%

Figura 3. Producción acuícola en la región y el mundo (miles de TM)
Fuente: FISHSTATJ (FAO). Elaboración de ComexPerú

De acuerdo a la FAO (2019), aunque el Perú ha tenido un crecimiento en la acuicultura de 1150 T en 1980 a 90 976 T en 2015, a un tasa de 2.23% en promedio de la producción anual anterior, a pesar de ello, Perú aún se encuentra muy por debajo de países líderes como Chile, Ecuador y Brasil para los cuales la producción nacional representa aproximadamente el 8.6%, 15.8% y 21.3% de la producción de estos países respectivamente. (Sociedad de Comercio Exterior del Perú, 2018)

La especie más representativa de esta actividad en el país es el langostino el cual se ha destacado por tener el mayor volumen de exportación de entre las especies producidas por esta actividad en el Perú.

1.2. Situación actual de la crianza de langostinos

Una de las especies más importantes en el crecimiento de la producción acuícola de especies para el consumo humano es el langostino, llamado en los países bajos como camarón o gamba dependiendo si es pequeño o grande, por este motivo es necesario realizar una visualización de la situación actual de la acuicultura de esta especie en el mundo y el Perú.

Según la FAO (2018), en el año 2016 se consiguió un nuevo máximo de capturas de las especies con las producciones más importantes, las langostas, gastrópodos, cangrejos y los

langostinos, cuyo valor promedio es 8 800 USD y 3 800 USD por tonelada. Como se observa en la Figura 4, a pesar de las oscilaciones anuales estas especies presentan una proyección ascendente en producción. Se destaca el crecimiento de la captura de los camarones, una manera alternativa de cómo llamar a los langostinos, el cual marca un ascenso más pronunciado que el resto de estas especies durante de 1994 al 2016, con un tamaño de producción máximo para el año 2016 de casi 3500 toneladas.

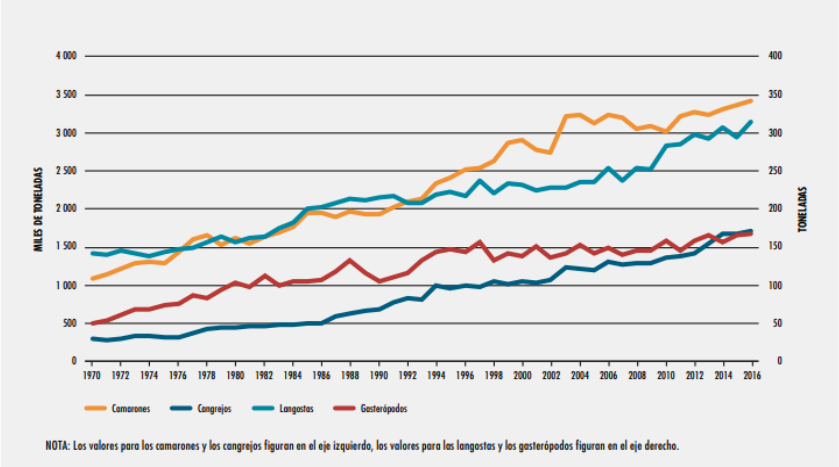


Figura 4. Tendencia en las capturas de grupo de especies valiosas
Fuente: FAO. El estado mundial de la pesca y la acuicultura (2018)

Los langostinos constituyen el segundo grupo principal en términos de valor en ser exportados. La alta comercialización de este producto a los mercados desarrollados como Estados Unidos y la Unión Europea, son abastecidos por la producción cuyo porcentaje más elevado se registra en América Latina y Asia Oriental y Asia Sudoriental. Aunque la captura de langostinos suministra grandes volúmenes, actualmente la mayoría del total del suministro total son cultivados. (FAO, El estado mundial de la pesca y la acuicultura, 2018)

Según FAO (2018), la China y Tailandia son los principales productores de langostinos cultivados, sin embargo, por retos como enfermedades y condiciones climatológicas generan la necesidad del crecimiento de la producción en otros países como Ecuador y la India. El aumento del interés por esta especie se debe al incremento de la demanda en países con economías emergentes cuyas preferencias evolucionan como consecuencia al aumento de sus ingresos. Además, ante este incremento de la demanda el precio comercial de este producto ha ido a aumento a lo largo de los últimos años.

Según Prom Perú (2017), las importaciones a nivel mundial de langostinos en el 2016 representaron aproximadamente 15 mil millones de dólares. Los principales destinos de las

exportaciones peruanas de langostinos son Estados Unidos, España, Francia, Canadá y Japón, representando ingresos de 89, 31, 10, 7 y 7 millones de dólares respectivamente. Perú se encuentra posicionado en el puesto décimo cuarto de países exportadores de langostinos. (Departamento de inteligencia de Mercados, 2017)

A pesar de la fuerte presencia de Chile en la acuicultura, existen otros protagonistas en cuanto a la producción de langostinos cultivados. Entre los países vecinos de América Latina y El Caribe se tiene como país líder a Ecuador con una producción de casi 194 628 toneladas, seguido por países como México y Brasil con una importante producción de 136 470 y 66 120 toneladas, también destacan países como Nicaragua, Colombia, Venezuela con producciones de 20 131, 18 639 y 16 763 Toneladas respectivamente. (Ministerio de la producción, 2017)

Según Produce[1], en el año del 2016 la cosecha nacional del langostino alcanzó 20 441 TM (toneladas métricas) representando el 20.7% del total de especies acuícolas producidas a nivel nacional. Este crecimiento en las exportaciones de langostinos es también producido por la aprobación del ingreso de langostino peruano a este mercado cuya demanda se valoriza en 300 millones de dólares (Departamento de inteligencia de Mercados, 2017).

La relevancia del langostino en la acuicultura se evidencia en las exportaciones de recursos hidrobiológicos provenientes de esta actividad, como se muestra en la Figura 5, para el año 2016 representó el 70.4%, es decir, es muy relevante su presencia en las exportaciones nacionales frente a especies como conchas de abanico, trucas, entre otras provenientes de la acuicultura. De acuerdo con cifras de la SUNAT, las exportaciones de langostinos registraron un total de 19 000 T durante el año 2017, alcanzando un crecimiento promedio anual del 13.7% en el periodo de 2013 y 2017.

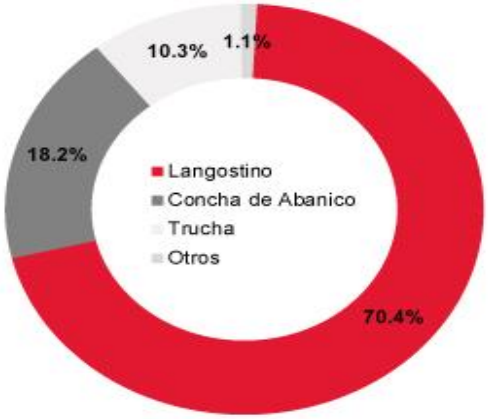


Figura 5. Exportaciones de recursos hidrobiológicos procedentes de la actividad acuicola en 2016
Fuente: Sunat. Elaboración: ComexPerú

[1] Ministerio de la producción

Para el periodo de enero-agosto del año 2018 se registró un valor por exportación acuícola de langostinos, conchas de abanico y truchas de 203 millones de dólares, lo que representa un incremento del 3% respecto a lo obtenido en el mismo periodo del 2017. Este crecimiento se explica por, como se muestra en la Figura 6, el incremento del 6% de las exportaciones del langostino en el 2018 respecto a las 2017 en este mismo periodo y el crecimiento de la exportación de la trucha. Esto evidencia el continuo crecimiento de los volúmenes exportados de langostinos cultivados en el territorio nacional. (ComexPerú, 2018)

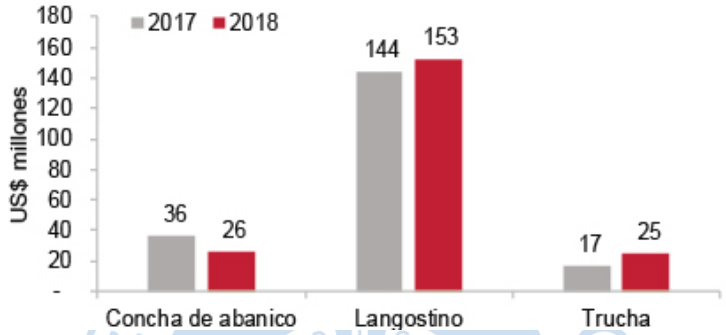


Figura 6. Principales exportaciones de productos acuícolas (enero-agosto)
Fuente: Sunat. Elaboración de ComexPerú

El consumo de langostinos tropicales ha aumentado sustancialmente en los últimos años en los Países bajos, Francia y Bélgica, se ha encontrado en un gran desarrollo acompañado por una disminución del precio, debido a la oferta misma. A pesar de esto, los langostinos son uno de los productos con mayor valor. Los mercados suelen estar internacionalizados por lo cual la demanda local suele ser abastecida por la importación de productos más baratos. Esto genera que haya un creciente interés por mejorar los procesos productivos de este producto debido a su alta demanda en los mercados y la alta competitividad de países productores acuícolas de langostinos.

1.3. Proceso productivo del cultivo de langostino en agua dulce

Existen diversas especies de langostinos de entre las cuales destaca, por su volumen de producción y presencia en el mercado, el *Penaeus vannamei* o *Litopenaeus vannamei.* En el Perú se cultiva esta especie conocida como "langostino blanco", es por este motivo que el proceso productivo que se describirá líneas abajo es del *Litopenaeus vannamei;* a partir de ahora cuando se mencione la palabra langostinos se hace referencia a esta variedad.

Esta variedad de langostino se caracteriza principalmente por la capacidad de crecer bajo confinamiento en altas densidades y reproducirse en cautiverio. En libertad llega a alcanzar un peso de 50 g y hasta 30 g en criaderos.

El proceso productivo se puede dividir en cuatro etapas, así como se muestra en la Figura 7. La primera de laboratorio, denominado "*Hatchery"*, donde se desarrollan actividades como selección, fecundación, desove, y desarrollo larvario y post-larvario. Otra fuente alternativa a los laboratorios para obtener estas semillas para cultivar es importándolas, es decir, que en este caso iniciaríamos con la etapa de cultivo directamente. La tercera etapa es la del procesamiento del langostino y finalmente el mercado de exportación. (Produce, 2008)

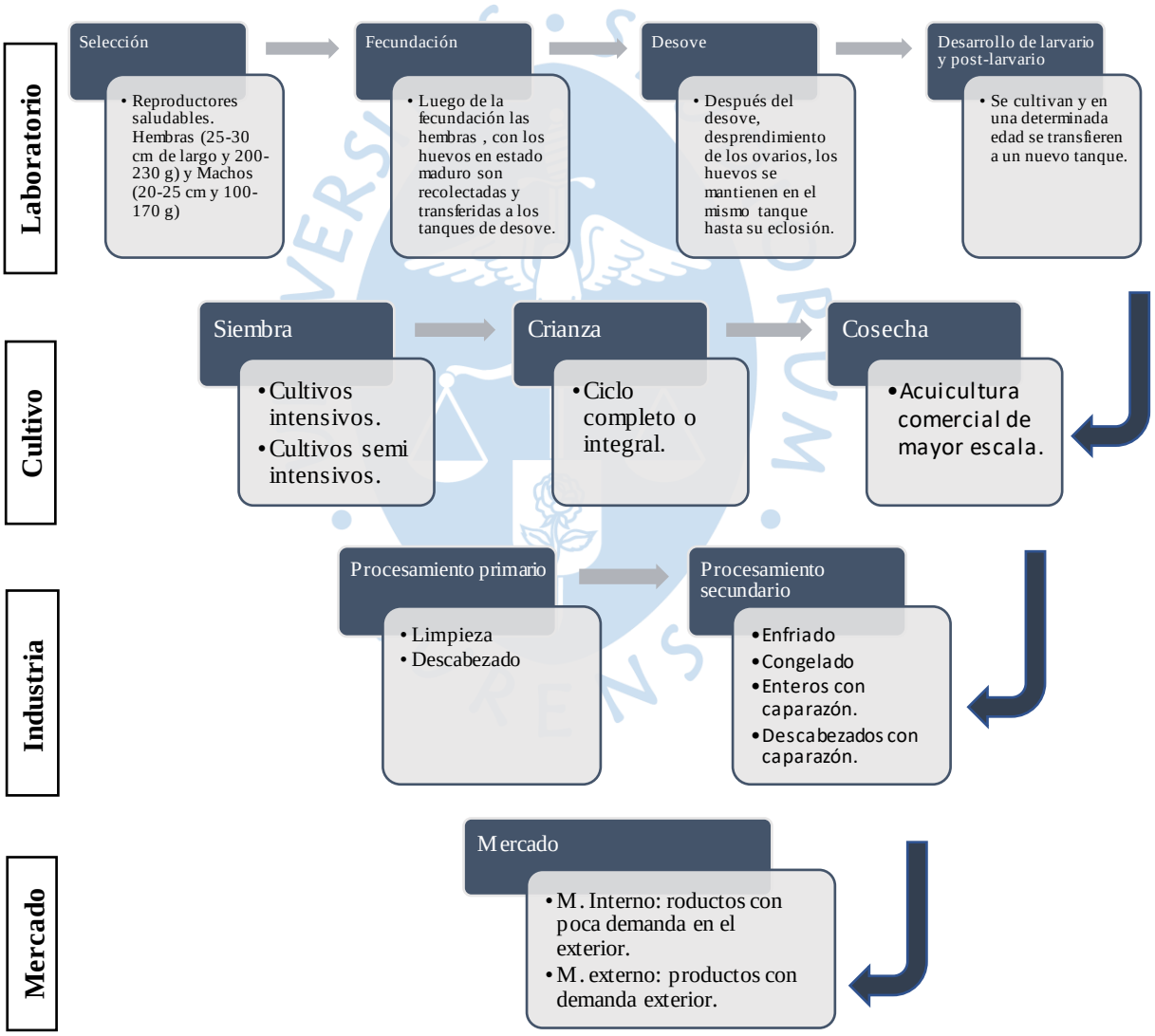


Figura 7. Cadena productiva del langostino
Fuente: tomado y adaptado de (Produce, 2008)

Para efecto de esta investigación se centrará en el proceso de cultivo del langostino. Esta etapa es la de mayor importancia debido a su mayor duración en comparación a las otras tres, y es donde se desarrollan las actividades claves de siembra, crianza y cosecha que son de principal interés para los productores.

El proceso de cultivo de langostinos en el Perú, según el Instituto del Mar del Perú (2002), cuenta con las siguientes generalidades técnicas y características:

- Infraestructura: el cultivo se desarrolla a un nivel semi-intensivo en ambientes controlados, se emplean estanques nivelados cuyas superficies varían entre de 5 a 10 ha. El tamaño de las granjas varía, según la empresa, entre 20 a 200 ha.
- Manejo de agua: el abastecimiento de agua es por bombeo con capacidades que varían entre 0.5 a 3 m^3/s. El agua es obtenida de esteros (termino usado para referirse a ambientes pantanosos) y se distribuyen por medio de un canal principal, reservorio, y canales secundarios.
- Densidades de siembra: la siembra se realiza de forma directa y eventualmente se hace uso de tanques de precria, con una densidad que varía entre los 10 y 30 ind/m^2 (individuos por m^2).
- Manejo de alimentos: se utiliza un alimento balanceado cuya aplicación recomendada, por sus ventajas técnicas, es por medio de "comedores".
- Monitoreo de la calidad del agua: actividad rutinaria de controlar la transparencia, oxígeno, salinidad, nutrientes (N, P) y elementos tóxicos en los estanques para adoptar medidas correctivas.
- Cosecha: la cosecha se realiza cuando los ejemplares alcanzan un peso entre los 12 a 18 g. Siguiendo las fases lunares, se hace uso de una bolsa en la compuerta y se procede al secado del estanque.
- Patología: el control sanitario de los organismos en cultivo se realiza con el fin de prevenir enfermedades. Los principales problemas son generados por bacterias (Vibrio, Pseudomonas, Rickettsia) y virus (TSV, IHHN, WSSV).

La crianza es una etapa previa a la siembra de las semillas donde se mantienen en tanques de calentamiento entre 1 a 5 semanas hasta que alcancen tallas entre 0.2 a 0.5 g. Para la siembra propiamente, las técnicas para el crecimiento se pueden dividir en cuatro categorías según la densidad de siembra. Las técnicas extensivas se refieren a aquellas siembras cuya densidad es baja, muy común en Latinoamérica, requiere estanques de con una superficie de entre 5 a 10 ha con densidades de entre 4 a 10 ind/m^2 (individuos por metro cuadrado). La semi-intensiva se

caracteriza por densidades de siembra media de entre 10 a 30 ind/m^2 y requieren estanques con superficies de entre 1 a 5 ha. La intensiva es una técnica con altas densidades de siembra de entre 60 a 300 ind/m^2 y con estanques con superficies de entre 0.1 a 1 ha. Finalmente, la técnica super-intensiva con una extremadamente alta densidad de siembra que puede ir desde 300 a 450 ind/m^2, con canales de aproximadamente 282 m^2.

Las cosechas de los estanques de cultivos extensivos y semi-intensivos se realizan a través de redes, con las que se busca acorralar a los langostinos a un lado del estanque, y aprovechando las mareas bajas de lo contrario debe bombearse el agua. Mientras que las cosechas de cultivos intensivos se realizan de manera similar. En la Figura 8 se muestra el ciclo productivo del langostino según la FAO (2009) que ya se ha descrito líneas arriba.

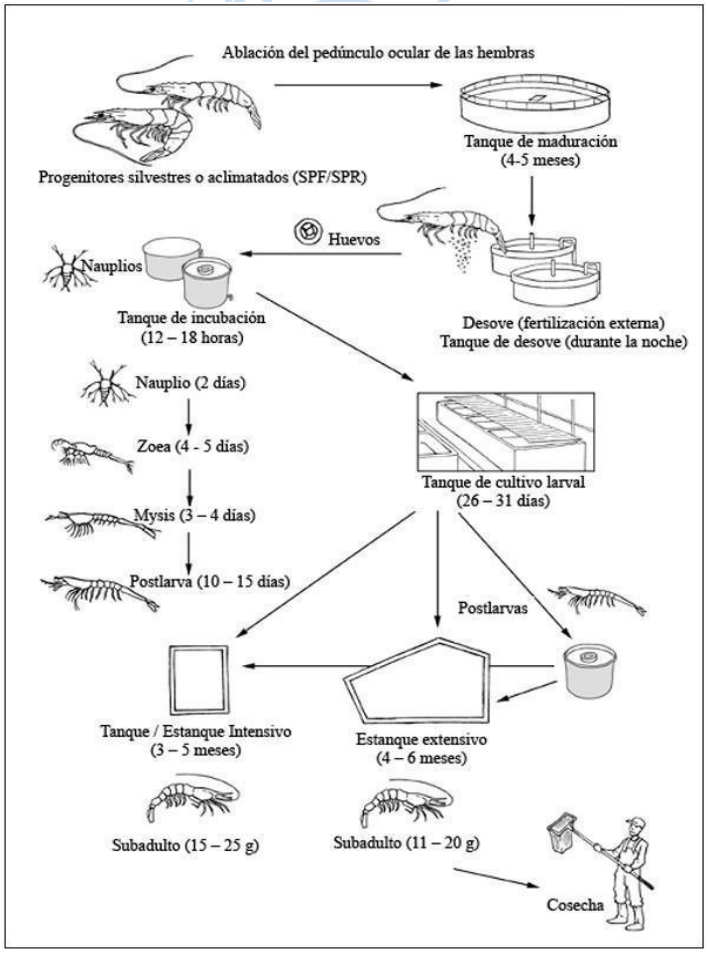


Figura 8. Ciclo de producción del langostino
Fuente: recuperado de FAO (2009)

Sin embargo, en actualidad el proceso nacional de crianza de langostinos se desarrolla basándose en conocimientos empíricos que se han adquirido a lo largo de la práctica. Por ello surge el interés por desarrollar los procesos de esta actividad de forma controlada y estudiada.

Una manera de cómo mejorar este proceso de cultivo de langostino, es el diseño de metodologías estadísticas avanzadas para el control y mejora de la productividad en procesos acuícolas. Determinar cuáles de las variables aportan o explican el crecimiento del langostino, expresado en peso, y de esta manera poder identificar qué valores y características de estas variables contribuyen a un proceso de crianza del langostino más controlado.

1.4. Proceso del monitoreo del crecimiento de langostinos en Perú y en el mundo

Frente a la emergente presencia de la acuicultura del langostino en el mundo, y en consecuencia en Latinoamérica y el Perú, se crea un interés en mejorar el proceso de producción de langostino de esta región del mundo, para aprovechar su potencial a través de mejoras en el proceso productivo. Para hacer esto es necesario reconocer aquellas variables que son importantes en el proceso y de esta manera saber qué medir, comparar para controlar y finalmente realizar las mejoras en el proceso.

El monitoreo, entonces, junto con las herramientas para realizarlo toman un mayor interés de investigación. El cual debe tener en consideración que el crecimiento del langostino es afectado por el tiempo, calidad del agua (Temperatura, salinidad, oxígeno disuelto, pH, compuestos de nitrógeno y sulfuro de hidrógeno), y el manejo de los estanque o piscinas (preparación del estanque, densidades de los estanques, prácticas de cultivo, técnicas de cosecha).

El manejo de los estanques es un estándar entre las empresas de cultivo de langostinos, salvo por ciertas particularidades, este proceso en general es el mismo, así que, no representa una diferencia sustancial entre empresas. Es por este motivo que el tiempo y la calidad del agua son los factores bajo los que se desarrollan gran parte de las investigaciones relacionadas al monitoreo de este proceso.

Investigaciones como las realizadas por Nguyen Tang, Nguyen Dinh, Tra Hoang, & Luong Hong (2015); Mishra, Verdegem, & Van Dam (2001); Carbajal Hernández, Sánchez-Fernández, & Villa Vargas (2012); etc. abarcan desde la implementación de sensores inalámbricos que se relacionen a sistemas integrados para mantener el control de la calidad agua, a través del análisis de estos datos, hasta la aplicación de herramientas de *data science,* como *machine learning,* que tienen usualmente como finalidad realizar predicciones o estimaciones del crecimiento del langostino, todas son técnicas utilizadas en el mundo que contribuyen al monitoreo de este. Existen también investigaciones locales como la realizada por León Caminati (2017) en el que se plantea el uso de herramientas estadísticas, en este caso, cartas de control para monitorear las variables críticas en el proceso.

Por lo mencionado anteriormente, la aplicación de herramientas estadísticas y de *machine learning* son un punto fundamental para el monitoreo del proceso productivo del langostino en el Perú y el mundo. Estas técnicas permiten entender mejor y analizar los datos para encontrar información relevante sobre el proceso y tomar decisiones, fundamentadas en estas, parar mejorar este proceso. Una de las aplicaciones más comunes en el mundo es el modelado de las curvas de crecimiento, las cuales se construyen bajo diferentes finalidades y es en función de esta misma que resulta más o menos virtuoso el uso de una por otra de las técnicas mencionadas anteriormente.

1.5. Empresa ECOSAC

ECOSAC AGRÍCOLA SAC, es una empresa agroexportadora dedicada a la producción y comercialización agrícola, acuícola, congelado de productos hidrobiológicos, procesado de conservas y empaquetado de fruta, que satisfacen la demanda de los clientes. Esta empresa facilita los datos del cultivo de langostinos para el desarrollo de este trabajo.

Se encuentra ubicada a 8 km la ciudad de Piura y a 70 km del puerto más importante del norte del Perú y cuenta con más de 1500 trabajadores fijos y 4500 operarios intermitentes. Al cierre de actividades del año 2018 la compañía contaba con un área disponible en campo de 6 000 ha de las cuales tiene ocupadas aproximadamente 2 276 ha donde se desarrollan las actividades agrícolas correspondientes a toda la variedad de productos de su portafolio. Para el riego se utiliza agua proveniente del río Piura que a través de canales se riega primero las zonas de cultivo de los langostinos. (Equilibrium, 2019)

Cuenta con 3 plantas una de frutas, conservas y otra de langostinos, esto debido a su línea de productos comprende palta, mango, uva de mesa, pimiento del piquillo, pimiento marrón, páprika y langostinos, nos centraremos de manera más específica en este último producto que ofrece la empresa. El cultivo de langostinos en la empresa inicia en el año 2000, se iniciaron pruebas en pozas que propiciaron un proceso de crecimiento y aprendizaje sobre este nuevo emprendimiento.

La compañía produce y exporta a través de Eco Acuícola S.A.C. langostinos enteros, cola de langostinos y langostinos con valor agregado. En el último año, la campaña de langostinos del 2019 que inicia desde enero hasta agosto de este mismo año tiene como destino China, se espera que el 70% de las exportaciones tengan como destino China aprovechando la comercialización de los langostinos enteros, y el 30% restante se destina al mercado de Estados Unidos.

Para el cultivo de los langostinos se tienen destinados aproximadamente 810 ha de pozas que producen aproximadamente 4 500 TM de este producto, el cual a partir de la campaña 2015 se procesa en la una propia planta a pocos minutos de las pozas. Específicamente las actividades de producción y procesamiento se realizan en un fundo ubicado en el caserío de Chapaira a 10 km de la Ciudad de Piura y a 1.5 km del Río Piura.

Según ECOSAC (2019), en cuanto a la exportación del langostino la empresa tuvo una participación en el año 2018 del 8.91% lo que representa un ingreso por exportaciones de 222.04 millones de dólares, convirtiendo a ECOSAC en la posición 4 en el sector. Esta actividad ha tomado un mayor interés por parte de la empresa a causa del ascenso del valor de los activos biológicos a 5.51 millones de dólares lo que representa un incremento del 50.76% respecto al año 2017, acompañado de la intensión por producir una mayor cantidad de langostino entero para China.

Este producto tiene diferentes presentaciones con las que se hace presente en el mercado mundial, estas son:

Tabla 1. Presentaciones del langostino

Productos	Descripción
Cola Shell-On	Descabezado congelado en bloque.
P&D tail on IQF	Pelado y devenado con cola, congelado rápido individual.
P&D tail off IQF	Pelado y devenado sin cola, congelado rápido individual.
BTO (Butterfly)	Cola pelada con corte longitudinal (corte profundo)
BTR (Butter round)	Cola pelada con corte longitudinal (corte medio)
EZ Peel	Cola con caparazón, corte longitudinal

Fuente: tomado y adaptado de ECOSAC (2015)

En la Figura 9 podemos observar que, en cuanto a composición de ventas por productos en la empresa, langostinos se ubican en el segundo puesto superado por la producción de uva y seguida por pimiento piquillo, pimiento marrón, entre otros menores.

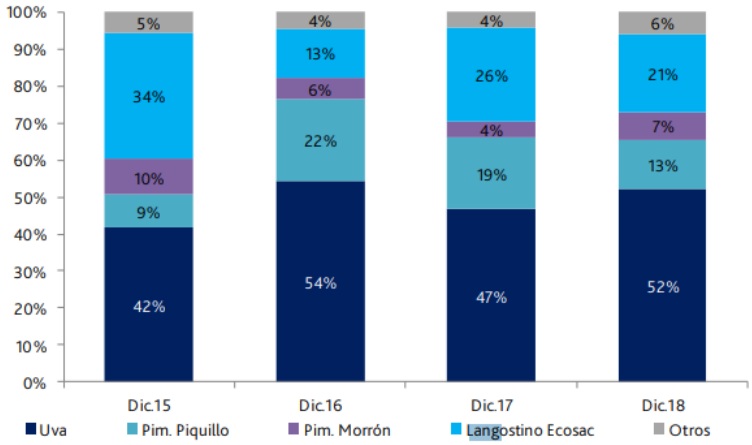


Figura 9. Composición de Ventas por Producto (%) de ECOSAC
Fuente: recuperado de Equilibrium (2019)

UNIVERSITAS STUDIORUM PIURENSIS

Capítulo 2

Análisis de datos

En este capítulo se describirá la condición inicial de los datos que se han utilizado para realizar el modelamiento, también el tratamiento de los datos, la identificación de datos atípicos que podrían generar alguna dificultad durante el análisis de la data por medio de las herramientas estadísticas y la depuración de estos mediante procedimiento continuo para asegurar la limpieza total de los datos.

2.1. Limpieza

El conjunto de datos que se tuvo a disposición, se encontraba almacenada en libros Excel, ordenados por siembra y campaña, cada hoja de estos libros representaba a cada una de las piscinas, y dentro de cada una de estas hojas se tienen datos generales de la piscina, como el número de la piscina, el área expresada en hectáreas, el peso de siembra, la cantidad de siembra, etc., y datos sobre parámetros e indicadores que se miden diariamente por colaboradores de la empresa. Para el primer análisis se utiliza data de la siembra 1, campaña 15 (2013-2014), así que esta data es tomada como base para realizar el modelamiento y la explicación de los capítulos, se planea realizar el mismo ejercicio con la data de las siguientes campañas.

Los datos, según la manera en la que se encuentran dispuestos, no permite que se les analice, debido a que es necesario que la información se disponga en una sola hoja que contenga la información relevante para el análisis, así que, por esta necesidad se realizó la adecuación de la data a un nuevo Excel en el cual se integra la información semanal de cada una de las piscinas por siembra.

Los indicadores como fecha, peso, total de animales, etc., son mediciones que se realizan a inicio de la semana que le corresponde, también existen variables que son el promedio de las mediciones diarias que se hacen como alimento promedio, biomasa promedio, temperatura, etc., y también datos que son los acumulados como kW totales. Entonces, por ello, es necesario

uniformizar la data de manera que todo se encuentre en una sola hoja de trabajo los valores promedios o acumulados por semana y por piscina.

2.2. Parámetro o indicadores

Para entender la naturaleza de los datos es indispensable comprender que representa cada uno de los parámetros o indicadores más importantes que se encuentran registrados, por ello a continuación se describirán de forma breve:

- **N° Semana:** el número de la semana desde que el langostino es plantado en una piscina.
- **Peso:** es el peso promedio en gramos del langostino por semana por piscina. Es la variable más importante debido a que es el principal indicador del crecimiento del langostino.
- **Biomasa promedio:** es suma de biomasa perteneciente a la evaluación de mortalidad diaria (EMD), evaluación poblacional (EP) y evaluación por consumo de alimento (ECA). Se calcula como el producto del total de animales de cada uno por su peso dividido entre 1000 para dimensionar a kg, dividido entre la cantidad de las 3 evaluaciones que se hayan registrado.
- **Total de animales:** según los documentos de registro[2] de la empresa ECOSAC, Microsoft Excel (2010), lo definen como la suma del total de animales correspondientes a la evaluación de mortalidad (EDM), evaluación poblacional (EP) y evaluación por consumo alimento (ECA), el primer indicador toma el valor "0" si la densidad ECA es mayor a "0", pero si es "0", según los documentos de registro, entonces:

$$Total\ animales\ EDM = cantidad\ de\ siembra \times \left(1 - ({}^{5}/_{160} \times número\ de\ días)/100\right)$$

Mientras el segundo y tercero se calculan como producto de su correspondiente densidad y el área de siembra.

- **Alimento:** es la cantidad de alimento que se le entrega a las piscinas, expresado en kg. Hay tres campos relacionados al alimento: el alimento acumulado, que es la suma acumulada de esa semana y de las anteriores; el alimento semanal, que es la cantidad de alimento entregado en la semana correspondiente; y alimento promedio, que es la cantidad de alimento promedio diario.
- **Factor de conversión semanal:** es la relación de la cantidad alimento y el incremento del peso en el total de animales. Es decir, su expresión sería la siguiente,

[2] Hoja de registro y almacenamiento diarios de indicadores y parámetros de la primera siembra de la campaña 15.

$$\frac{Alimento\ semanal}{(Total\ de\ animales * imcremento)/1000}$$

- **Factor de conversión acumulativo:** es la relación entre la cantidad de alimento acumulado y la biomasa promedio, es decir, indica cuanto del alimento acumulado se transforma en biomasa.
- **Nicovita, Gisis y Purina:** es la cantidad de cada tipo de alimento que se utiliza por día, existen diferentes presentaciones que son registradas y almacenadas. La nicovita es una marca de alimento para animales. En el caso de los langostinos tiene diferentes presentaciones según la finalidad, por ejemplo, aquellas concentradas en la digestibilidad y estabilidad del alimento en el agua, crecimiento acelerado y reducción del tiempo de cultivo, sobrevivencia al inicio de la producción, etc. (Nicovita, 2018).

  El alimento de gisis, es producida por la marca ecuatoriana Gisis S.A. dedicada a la producción de alimentos balanceados para animales, entre ellos los langostinos, para estos ofrece un nutrición basada en presentaciones clasificadas por tamaño del langostino, entre estas resaltan las de la línea intensiva "I", que es diseñada para las fases juvenil y adulto, según el trabajo de investigación por la ingeniera S. Caballero Solano (2015).

  La purina, según Díaz Herrera, Juárez Castro, Pérez Cruz, & Bückle Ramírez (1991), es un tipo de alimento que se aplica por lo general en la etapa de postlarva o juvenil del langostino debido a que gracias a su contenido de proteínas, lípidos y carbohidratos de calidad, que son indispensables para el desarrollo de los langostinos. También hay "otros" que representa la cantidad de alimento cuya marca no está incluida en las anteriores, sin embargo, hay registro de la cantidad y el día del uso de este tipo de alimento.
- **Aplicaciones:** en este se incluyen aplicaciones adicionales que se les dan a las piscinas con el fin de mejorar la estabilidad del alimento, adecuar el alimento, etc., entre los que se registran podemos destacar el hidróxido de calcio, sal, vitamina C, entre otros.
- **Aireación:** la cantidad de kW que se entrega a la piscina para oxigenar el agua, esto con el fin de contrarrestar el consumo de oxígeno que se genera por la biomasa presente en las piscinas. Se registran la cantidad de aireadores y el número de horas que se encuentran trabajando, en consecuencia, se puede calcular el 'Total de kW'.
- **Parámetros:** se registran diariamente los valores máximos y mínimos de tres parámetros. La temperatura, que es la variable que tiene un mayor impacto en el desarrollo en los procesos químicos y biológicos, incluso se considera que esta variable guarda una relación directamente proporcional con el crecimiento del langostino, según Ulloa Tello R. (2015).

El oxígeno, es el parámetro que guarda una relación con el crecimiento de materia orgánica, debido que, a consecuencia del crecimiento de los langostinos, crece también la demanda de oxígeno, esto se ve reflejando la cantidad de demanda de oxígeno disuelto que hay en las piscinas, según Ulloa Tello R. (2015).

El pH, es un indicador del estado medio, si es ácido o básico, se suele relacionar a las piscinas acuícolas que tienen pH bajos con tasa bajas de crecimiento, reproducción o supervivencia, se suele agregar cal agrícola para regular los pH bajos, según Ulloa Tello R. (2015). Se destacan estos parámetros debido a que son aquellos que pueden describir las condiciones bajo las cuales el langostino se desarrollan.

- **Salinidad:** se registran la cantidad (ppm) de sales minerales que se encuentran disueltas en las piscinas, su valor depende de básicamente siete iones sodio, magnesio, calcio, potasio, cloruro, sulfato y bicarbonato. Se encuentra fuertemente relacionado con las temporadas de lluvia, debido que, cuando las hay los niveles de salinidad disminuyen y aumentan cuando hay tiempos de sequía. En la data se tiene pocos registros, lo cual dificulta que se considere como un factor que afecte al crecimiento del langostino.
- **Química del agua:** en esta clase de parámetros, se registran los valores diarios de indicadores como alcalinidad (CaCO3), amonio (NH3), nitrito (NO2), nitrato (NO3) y fosfato (PO4), los cuales buscan describir la química del agua de las piscinas en las que se desarrolla el langostino.
- **Productividad columna de H2O:** se registran la cantidad de algas que aparecen en las piscinas por la presencia de materia orgánica. Estas algas son las cianofitas, que es un tipo de alga verde azulada que se le asocia a las bacterias debido a la ausencia de membrana, son conocidas por su estabilización de suelos para evitar la "voladura"[3], así mismo también son usadas como fertilizante.

  Sin embargo, su presencia en grandes cantidades puede alterar el color, sabor y olor del agua, esto es muy probable debido a que con los nutrientes necesario y bajo las condiciones de luz y temperatura este tipo de alga puede multiplicarse fácilmente, aunque no suelen representar un problema mayor existen casos que evidencia que contienen cuerpo que pueden desprender toxinas nocivas para los animales, estas suelen ser *Anabaena spiroide*s, *A. Circinalis* y *A. Lemmermanni*, según Echenique R. & González D. (1998).

  Las clorofitas o más conocidas como algas verdes cuyo tamaño puede variar entre magnitudes microscópicas hasta aquellas que son formadas por filamentos de longitud

[3] Termino para describir un suelo el cual no puede retener la humedad por un periodo prolongado.

considerable, de las cuales el 90% de esta especie de algas son de agua dulce y el otro 10% son de aguas saladas, según Guevara A. & Calix L. (2015).

Las diatomeas, según Guzman B. & Leiva D. (2015), son algas microscópicas, unicelulares y eucariotas, que son altamente sensibles a los cambios de la composición de los nutrientes (P, N, Ni), y la composición de comunidades de diatomeas implica de forma indirecta la calidad biológica del agua.

Las euglenofitas, según Murray Nabors (2011), es un tipo de algas que generalmente se encuentran en las aguas dulces, suelen ser incoloros y los que cuentan con pigmentación se aproximan eventualmente a las incoloras, todas poseen la capacidad de absorber moléculas orgánicas.

En consecuencia, la data que puede ser utilizada para el análisis, es aquella a la cual le corresponde un valor de peso, para de esta manera poder encontrar la relación con las demás variables que se presuponen tienen efecto ellas, en caso contrario esas mediciones no se incluirán.

2.3. Conflictos en la data

Los conflictos que puede presentar la data son problemas que se generan cuando se reúne información de diferentes fuentes y realizando proceso no estandarizados de levantamiento de información. Dependiendo de la naturaleza de estos conflictos de data, se pueden clasificar, según Iván Amón Uribe (2010), de la siguiente manera:

a) **Datos duplicados:** este tipo de conflicto se regenera cuando una misma entidad del sistema real es representada por dos o más veces en una o varias bases de datos, es decir, no hay un identificador único. Esto sucede cuando al ingresar la información sobre dicha entidad se le identificad de una forma determinada, sin embargo, luego cuando se vuelve a ingresar nueva información de la misma entidad se le refiere de una manera distinta a la anterior, y este error se pude seguir produciendo "n" veces, generando en consecuencia, que se tengan "n+1" datos que aparentemente son diferentes entre sí, sin embargo, hacen referencia a la misma entidad. (Amón Uribe, 2010)

b) **Valores faltantes:** es el tipo de problema que hace referencia a la ausencia de valores en un atributo requerido, es decir, que si la data presenta este problema en referencia a valores no obligatorios para su análisis no habría inconveniente, por ello no todos los campos vacíos representan un problema. Medina y Galván, asistente regional y asistente de investigación de Unidad de Estadísticas Sociales (CEPAL), respectivamente, advierten sobre las implicaciones estadísticas cuando se realizan ingresos de datos o sustitución de información: "*La aplicación de procedimientos inapropiados de sustitución de*

información introduce sesgos y reduce el poder explicativo de los métodos estadísticos, le resta eficiencia a la fase de inferencia y puede incluso invalidar las conclusiones del estudio" (Medina & Galván, 2007, págs. 26-27). Ante este problema surgen formas de trabajar como la *Listwise Deletion* (LD), que es trabajar con data completa, Avaible *case* (AC) o *Pairwise Deletion* (PD), que es trabajar con información disponible, y los métodos alternativos como la imputación de datos, que es estimar los valores faltantes a través de alguna técnica.

c) **Valores atípicos:** según Chandola et. al. (2009), el término valores atípicos o *outliers* hace referencia a aquellos datos que no se adecuan a un comportamiento esperado. También referidos a ellos como datos atípicos, anomalías, observaciones discordantes, daños, contaminantes, peculiares, etc. Estos valores son extremadamente dañinos para los estimadores estadísticos, es por este motivo que se genera la importancia de aplicar alguna técnica para detectar estos valores y eliminarlos o sustituirlos.

La data con la cual se está trabajando cuenta con dos de los problemas mencionados anteriormente, el primero es de los valores faltantes en la data. Estos campos no medidos generan un problema al momento de ingresar una variable al modelo, debido a que, la cantidad registrada no es representativa en comparación al tamaño de mediciones de las demás variables. Además, no habría correspondencia de los valores de las demás variables con en estos campos vacíos.

El segundo problema son los valores atípicos, que representan un peligro en cuanto a la distorsión de los indicadores estadísticos lo cual derivaría en una incorrecta interpretación del análisis, porque si se les consideran a estas variables dentro del modelo y se realiza la interpretación del efecto de cada una de las variables, se pueden establecer relaciones erradas entre las variables y obtener indicadores con poca credibilidad y representatividad.

Entonces, integrada toda la data se inicia el análisis de *outliers*, para evitar usar en el análisis, variables cuyas medidas se encuentren en magnitudes muy grandes o pequeñas, es decir, valores extremos.

2.4. Análisis de los *outliers*

Dado que la data registrada por los colaboradores en campo y, posteriormente, digitalizados a una hoja de registro diario que se tiene para cada una de las piscinas, la posibilidad de cometer errores humanos es alta. Pueden existir errores de transcripción que podrían generar complicaciones para el análisis de esta e incluso se podrían determinar recomendaciones y conclusiones erradas, que a su vez generarían decisiones equívocas sobre el sistema generando pérdidas de recursos, tiempo y credibilidad.

Es por ello es de vital importancia que lo datos se encuentren libres de errores, en consecuente, es imprescindible comprobar si los valores con los que se trabaja son los correctos y no hay ninguno fuera de lugar, es decir, surge la necesidad de limpiar la data. Para ello se debe identificar aquellos datos que indican un posible error a causa de su magnitud o porque son incoherentes en comparación al resto de valores del mismo parámetro, esto se puedo lograr con el uso de alguna técnica para detección de atípicos, una vez identificados estos datos se deberá buscar la causa raíz de este comportamiento anormal corregirlo o retirarlo según sea su condición y por último comprobar.

2.4.1. Clasificación de *outliers*

Como parte inicial para el tratamiento de *outliers*, se debe reconocer la clasificación de *outliers* son lo que se están tratando, para de esta manera decidir el tipo de técnica para corregirlos, que más se adecue, y también para identificar su posible origen y proponer posibles medidas para mitigar el riesgo de generar atípicos en la data para futuros análisis de la data.

Los *outliers* pueden ser clasificados en dos tipos, según White R. A. (1992):

- **Univariados:** consiste en aquellos datos individuales que solo presentan valores extremos inusuales.
- **Multivariados:** son aquellos valores que resultan de la combinación de valores inusuales en al menos 2 variables. Este tipo de datos atípicos, también se pueden agrupar en dos, *gross outliers*, que son aquellos datos atípicos que son generados por valores individuales de las variables, y *structural outliers*, que son los valores atípicos generados por la estructura de la covarianza de aquellos valores que no lo son considerados como *outliers*.

Según Konrr (2002), dependiendo del comportamiento subyacente de los datos, los valores *outliers* pueden clasificarse de la siguiente manera:

- **Un valor extremo o relativamente extremo:** es aquel valor inesperado muy pequeño o grande, que no se adecua al comportamiento de la mayoría de los datos de la misma base de datos.
- **Un contaminante:** es aquella observación que, debido a la diferencia en el comportamiento con respecto al resto de datos, sigue de otra (posiblemente desconocida) distribución. Estos contaminantes pueden ser reconocidos o no, como *outliers*, dependiendo de si se ajusta a la misma distribución que siguen los demás datos.
- **Un valor legítimo pero sorprendente/inesperado:** es un suceso verdadero que no se encuentra contenido en el intervalo de valores del resto de la data, este tipo de *outlier* no

deben ser, necesariamente, eliminados sino dispuestos en las manos de expertos en el tema para que interpreten y den una repuesta de su origen.

- **Un valor que ha sido medido o grabado incorrectamente:** es aquel valor cuyo comportamiento anormal se justifica a causa de un error en la medición o grabado del dato, entonces, es necesario que sea corregido o eliminado de la base de datos.

En el caso particular de la data usada en este estudio, se encuentran *outliers* del tipo de univariados y multivariados, y en cuanto a la clasificación según Knorr se encuentran *outliers* de los cuatro tipos.

2.4.2. Técnicas para identificar *outliers*

Una vez identificados la clasificación de *outliers*, es necesario establecer la técnica de identificación de *outliers*. A continuación, se describen brevemente algunas de las técnicas de detección, según Adrián de Armas (2015):

- **Análisis de valores extremos:** análisis básico de detección de *outliers* para datos de una dimensión, en el cual se asume que los datos de valores muy grandes o pequeños son atípicos. Este análisis considera la definición de *ouliers* de Hawkins (1980), quien dice que los outliers se definen por la probabilidad de aparición, y no por su pertenencia a los extremos. Para dicho análisis, generalmente, se supone que un conjunto de observaciones siguen una distribución normal $N(\mu, \sigma^2)$, luego el análisis recae en reconocer aquellos datos que se encuentran en la "región *outlier*", según Ben-Gal (2005), que es representado por "α", donde $0<\alpha<1$, para ello es necesario recurrir a la distribución de la normal estandarizada y evaluar si el valor cae en la región de atípicos.
- **Modelos estadísticos y probabilísticos:** se modelan los datos a una determinada distribución de probabilidades, donde los parámetros del modelo se obtienen a partir de la data procesada, como los modelos trabajan con probabilidades es necesario la normalización para su análisis. Tiene la desventaja, que no siempre los modelos estadísticos se ajustan de forma adecuada a los datos, por ello es vital conocer bien el proceso o sistema que se desea modelar a través de las distribuciones de probabilidades, y asegurar que este sí representa la realidad. (De Armas, 2015)
- **Métodos basados en profundidad:** es un método gráfico que consiste en ordenar los datos por capas, envolventes convexas, con la suposición que los valores que se encuentran en la capa más superficial son valores atípicos, o al menos es improbable que se encuentren atípicos en las capas más profundas, este método permite evitar el problema de ajustar una

distribución de probabilidades a los datos, además de ser indicado para la detección de *outliers* cuando no es razonable utilizar la función de distancias métricas (Knorr, 2002).

- **Métodos basados en desvío:** es el método que se basa analizar la varianza de un conjunto de datos y evaluar el impacto de los *outliers* en ella, es decir, se buscan aquellos valores que incrementan la varianza, evaluando cómo varía esta cuando aquellos valores son eliminados, identificando de esta manera aquellos valores que generan una reducción significativa en la varianza cuando son eliminados. Luego de eliminarse los datos, se dice que el conjunto de datos se "suaviza", por esta razón se define un indicador llamado "*smoothing factor"* (SF), que indica la reducción de la varianza, del conjunto de datos, cuando se eliminan unos puntos de la muestra original (posibles anomalías). (De Armas, 2015)
- **Métodos basados en distancia:** es un tipo de algoritmo usado para la detección de atípicos a través del reconocimiento de la definición de un puntaje según la distancia que tenga este punto con respecto a los datos vecinos más cercanos, no requiere la construcción de un modelo para ajustar a los datos, solo trabaja con los vecinos más cercanos. Entonces, se consideran valores anormales a aquellos que tienen una distancia mayor en comparación a sus "k" valores vecinos. (Aggarwal, 2013)
- **Modelos lineales:** son herramientas útiles para la detección de valores atípicos, debido a que muchos conjuntos de datos suelen presentan correlaciones significativas, esta correlación es causada porque los atributos son generados por el mismo sistema o proceso, entonces genera una estrecha relación entre ellos.

  Existen dos clases de modelos lineales el primero es una regresión estadística para relacionar variables independientes con variables dependientes para definir dependencias entre ellas, y la segunda usa un análisis de componentes de manera que le da igual importancia a cada una de las variables para determinar los subespacios dimensionales inferiores de proyección, ambas son muy parecidas, pero difieren en la formulación de la función objetivo. (De Armas, 2015)
- **Modelos basados en proximidad:** es un análisis de reconocimiento del entorno de los puntos, evaluando la población alrededor de ellos, decir, si la población es escasa del punto, este es considerado como un *outlier*. Existen varias formas en cómo definir la proximidad en un análisis de valores atípicos, según Aggarwal (2013), pueden ser basado en *clusters*, es una técnica estadística multivariable de clasificación de datos a través de una tabla de casos-variables que tratan casos individuales en grupos homogéneos según la esencia de

los datos, basado en la distancia, determinando la proximidad entre los datos vecinos basándose en la distancia entre los datos más cercanos, y basado en la densidad, busca definir la densidad local de un grupo de datos que es la cantidad de puntos dentro de una región local y en función de estos valores locales se establecen puntajes para los *outliers*. (De Armas, 2015)

A partir de los datos atípicos identificados en la base de datos se debe determinar las medidas correctivas que corresponda para depurarlos para poder realizar un análisis en fundamentados a datos "limpios".

2.4.3. Medidas de depuración

Para continuar con el proceso de limpieza de la data, es importante mencionar y explicar cómo se gestionarán las anomalías en los datos, dependiendo del tipo de *outlier* del que se trate. Se han considerado las siguientes medidas correctivas:

- **Corregir los datos:** se corrigen aquellos valores irregulares, según cómo se comporten los valores vecinos más cercanos, esto claramente, es muy probable cuando se sospecha que hay errores por grabado de los datos o de medición, y los valores extremos que presentan difieren en gran medida con el resto de los datos.
- **Verificar la veracidad de los datos:** cuando los valores extremos no se alejen de forma considerable con respecto a los demás datos, es muy poco probable que provengan de errores de grabado y en consecuencia se supone que se trate de un valor real. Entonces, es necesario que se verifique la veracidad de los valores con los expertos del tema, es decir, con las personas que conocen mejor el sistema y pueden validar aquellos valores aparentemente atípicos.
- **No considerar los datos:** es una medida que se considera, en caso una variable cuente con campos vacíos, es decir, se evitará el uso de alguna medida de imputación de datos para evitar algún efecto de sesgado. Además, que se considera esta medida a aquellas variables que tienen pocos valores disponibles para su análisis, lo cual reduce la confiabilidad de las observaciones de esta variable.
- **Eliminar los datos:** es el tipo de medida que se evalúa tomar, cuando no se le encuentra alguna explicación del origen de este valor, tras la evaluación de la relación con los demás datos y previa consulta con los expertos, se considera la eliminación de este tipo de datos por considerárseles anomalías inexplicables y no razonables.

2.5. Procedimiento

El procedimiento de limpieza de los datos inició desde la integración de toda la data en una hoja de trabajo, luego se presumieron los tipos de *outliers* que serían más probables de encontrar, por error en grabado, valores contaminantes y valores extremos. Se analizaron posibles patrones de comportamiento de correlación entre las variables a través de las gráficas de matriz, como se muestra en la Figura 10, y mediante un modelo lineal que se describirá en el capítulo 3. Este modelo lineal, que corresponde a la etapa de modelamiento de los residuos, también contribuyó a la detección de valores atípicos. En consecuencia, la actividad de limpieza de la data fue un método iterativo entre esta y el modelamiento, teniendo en consideración lo anteriormente mencionado en cuanto a evaluación de *outliers*.

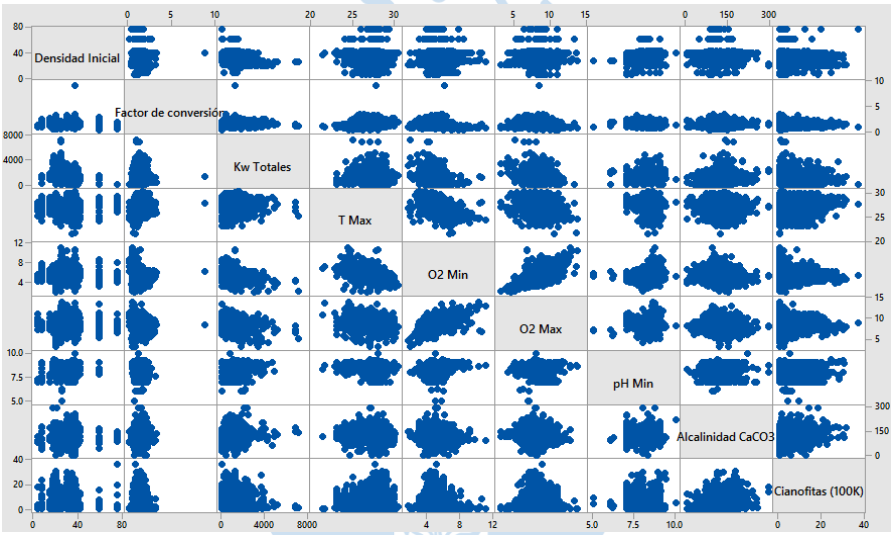


Figura 10. Gráfica de matriz de dispersión de Densidad inicial, Factor de conversión, etc.
Fuente: elaboración propia con información recuperada de la empresa ECOSAC.

A continuación, se mostrará un ejemplo aplicando el procedimiento descrito. En la Figura 11 se muestra una gráfica de matriz de dispersión entre las variables independientes "Factor de conversión" y "Densidad inicial". Se puede observar un aparente valor atípico del tipo "valor extremo" ya que la mayoría de los valores correspondientes a la variable "Factor de conversión" se concentran entre cero y cuatro, mientras que el éste es superior a 8, más del doble del valor máximo de esta variable. Probablemente se trate de un *outlier* del tipo de "valor grabado incorrectamente", debido a la naturaleza de la data.

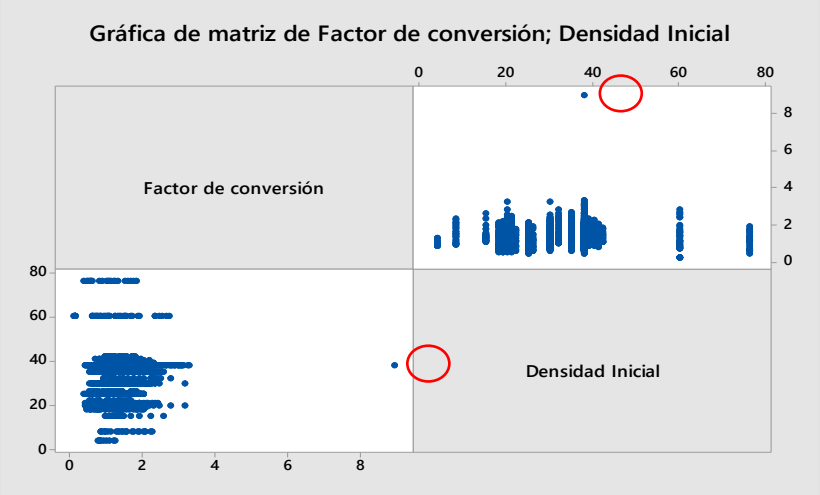


Figura 11. Gráfica de matriz de dispersión de Factor de conversión y Densidad inicial
Fuente: elaboración propia con información recuperada de la empresa ECOSAC.

Luego de Identificarse el dato anormal es necesario reconocer su origen, por ello es necesario, mediante la comprobación de su veracidad, buscar la causa raíz. Entonces, se encuentra que el dato se ubica en la fila 3726, con un factor de conversión de 8.91786, es decir, se trata de los datos de la semana 27 que corresponden a la piscina "P(5-47)".

Tras consultar a la data integrada se obtuvieron los datos mostrados en la Figura 12, de donde se puede concluir que el error proviene de un valor anormal en el "Total Animal", ya que sus valores antecesores se mantienen constantes en un valor y luego tiene un valor dramáticamente pequeño, de 370000 animales a 70000 animales, y esto asu vez afecta al "Promedio de biomasa", cómo se explicó subcapítulo 0.

Promedio de biomasa	Factor de conversión	Total Animal
9298.1	1.630763274	370000
9620	1.688773389	370000
9779.1	1.722142119	370000
9990	1.700800801	370000
10138	1.696685737	370000
1960	8.92	70000
535.31	0.799536717	378575
778.15	0.773629763	377744

Figura 12. Datos de la data integrada.
Fuente: elaboración propia con información recuperada de la empresa ECOSAC.

Pero como estos datos provienen de los documentos de registros, es imprescindible revisar la data original, como se puede observar en la Figura 13, los datos coinciden con los datos que se encuentran en la hoja de datos integrados, entonces error se origina a partir del registro en estos documentos.

Promedio Biomasa	Alimento			Factor Conversion Acumulativo	Factor Conversion Ideal	Factor Conversion Semanal	Total Animal	Total Alimento US$
	Acumulado	Semenal	Promedio					
9779.10	16,841	595	85	1.72	1.62	3.74	370,000	20,331.39
9990.00	16,991	150	21	1.70	1.62	0.71	370,000	20,512.12
10138.00	17,201	210	30	1.70	1.62	1.42	370,000	20,765.13
1960.00	17,479	278	40	8.92	1.62	6.62	70,000	21,100.08
#¡DIV/0!	17,686	207	30	#¡DIV/0!	#¡DIV/0!	#¡DIV/0!	70,000	21,349.48

Figura 13. Datos originales de los documentos de registro.
Fuente: elaboración propia con información recuperada de la empresa ECOSAC.

El comportamiento normal de la variable "Total Animal", en esta piscina, es de 370000 y no 70000. Entonces, debido a que es inconsistente un cambio tan radical del tamaño del cultivo de los langostinos se requiere que se realice una corrección a este valor y de esta manera corregir los valores de los demás indicadores que dependen de él.

Hecha esta corrección se obtiene valores mas adecuados y correspondientes a la tendencia que muestran los valores antecesores a este, como se muestra en la Figura 14.

Promedio de biomasa	Factor de conversión	Total Animal
8375.32	1.413677328	370000
8746.8	1.468422737	370000
8991	1.556445334	370000
9298.1	1.630763274	370000
9620	1.688773389	370000
9779.1	1.722142119	370000
9990	1.700800801	370000
10138	1.696685737	370000
10360	1.69	370000
535.31	0.799536717	378575
778.15	0.773629763	377744

Figura 14. Datos de la data integrada después de la corrección.
Fuente: elaboración propia con información recuperada de la empresa ECOSAC

En la gráfica que se muestra en la Figura 15, ya no se observa datos con valores tan extremos como en el anterior, así se podría repetir de manera sucesiva el identificar y corregir los valores atípicos hasta que no se encuentren evidentes valores atípicos por corregir.

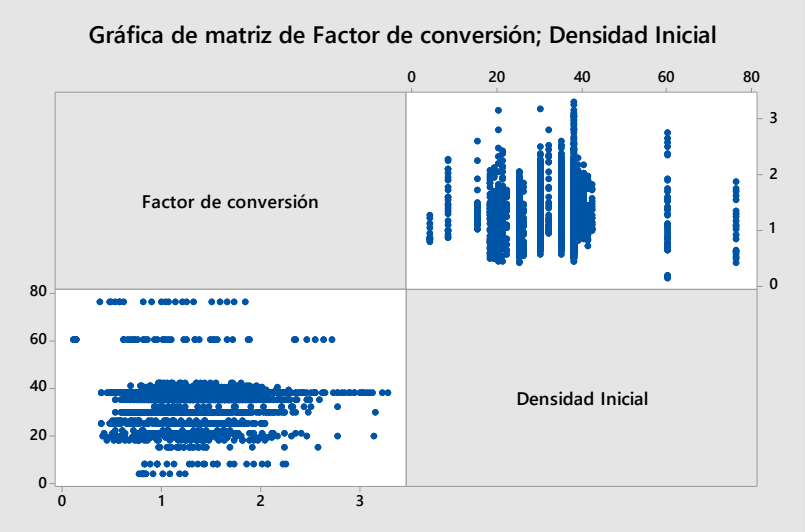


Figura 15. Gráfica de matriz de dispersión de Factor de conversión y densidad inicial después de la corrección
Fuente: elaboración propia con información recuperada de la empresa ECOSAC.

Después de realizar continuamente esta actividad de limpieza se consigue una base de datos confiable, a partir de la cual se realiza el modelamiento, que se desarrolla en el siguiente capítulo. En el Anexo A se muestra la media, deviación y varianza del peso de los langostinos por semana de cultivo, en base de la data limpia, y en el Anexo B y Anexo C se muestran las tablas con los valores medios y varianzas, respectivamente, de algunas variables de las primeras 10 piscinas.

Capítulo 3

Modelamiento y resultados

En este capítulo se describe la importancia del modelamiento, los métodos de modelamiento, la elección y descripción del modelo a utilizar. También se describe el procedimiento que se ha realizado para modelar el crecimiento del langostino, es decir, las consideraciones que se tuvieron en cuenta, supuestos y el modelo final, así mimo, se presentará los resultados. Este modelo final, construido con datos de la primera siembra de la campaña 15, permitirá establecer la relación entre las variables independientes y la variable dependiente, para que se pueda comprender y generalizar a data futura.

Consideraciones

- Se trabajará con data de la empresa, descrita anteriormente, dedicada a la siembra, cultivo y cosecha de langostinos en la región Piura, la cual contiene datos levantados por sus colaboradores durante las respectivas campañas, para un primer análisis se utilizó 5891 datos provenientes de la primera siembra de la campaña 15.
- La finalidad del modelamiento es la inferencia que se pueda realizar sobre este y no la predicción, así pues, se pondrá mayor rigurosidad a la interpretabilidad que a la exactitud del modelo.
- Para iniciar con el modelo es necesario reconocer el comportamiento del desarrollo del peso en los langostinos, el cual tiene una forma de "S" alargada, como se muestra en la Figura 16, que indica que tiene un crecimiento progresivo más acelerado cuando se encuentra en medio de su ciclo de vida, pero con la transformación del logaritmo natural adecuar la data, para que de esta manera sea más sencillo de modelar con una función.

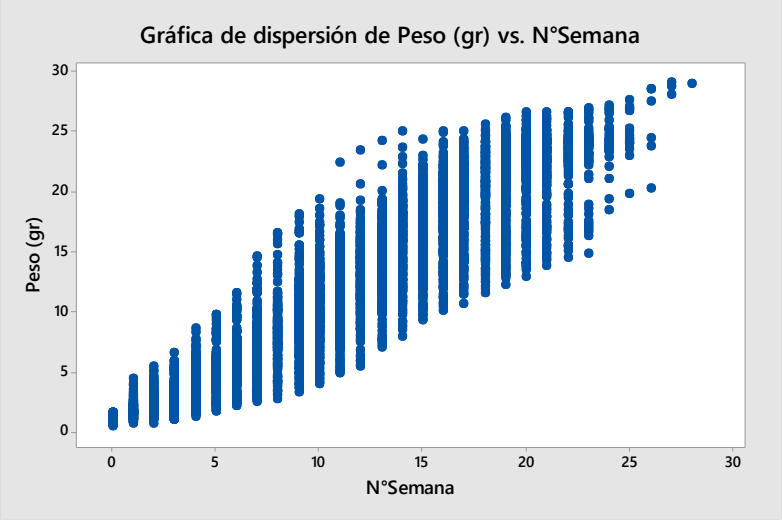


Figura 16. Gráfica de dispersión peso vs. N° Semana
Fuente: elaboración propia con información recuperada de la empresa ECOSAC.

Objetivo

- Estimar un modelo que permita describir el comportamiento promedio del crecimiento de los langostinos.
- Identificar a los factores más importantes o aquellos que tengan un impacto, estadísticamente hablando, más significativo sobre en cuánto difiere respecto al comportamiento medio del crecimiento de los langostinos.
- Identificar un modelo representativo que permita interpretar el efecto, de cada una de las variables, al crecimiento del langostino e interpretar la relación de cada una de las variables independientes que resulten significativas con la variable dependiente.
- Medir el impacto del oxígeno y temperatura que se sabe, por la experiencia del personal de la empresa, influyen en el crecimiento del langostino.
- Identificar cómo el tipo de alimento contribuye al crecimiento del langostino, si es significativo el tipo de alimento para el crecimiento o es indistinto.

Limitaciones

- Se determinará un modelo con el fin de entender el efecto de las variables con respecto al peso del langostino, es decir, no es finalidad el encontrar un modelo de predicción.
- Los datos y las variables son limitadas a las medidas por la empresa en las diferentes siembras de campañas pasadas.
- La data bajo la cual se desarrollará esta investigación corresponde al crecimiento del langostino de la variedad *Penaeus vannamei*, criada en agua dulce en la región Piura, por

lo tanto, el modelo se adecua mejor a esta especie y a condiciones en regiones con las características similares esta.

Hipótesis

- La temperatura guarda una relación directamente proporcional con el crecimiento del langostino.
- La temperatura y el oxígeno guardan una relación inversamente proporcional entre ellas.
- Los meses de siembra tienen un efecto en el crecimiento del langostino.
- Las densidades iniciales de la siembra de los langostinos afectan el crecimiento del langostino.
- La calidad y tipo del alimento afecta al crecimiento del langostino.
- El efecto del alimento acumulado pude afectar el desarrollo del langostino, es decir, el exceso de alimento acumulado puede generar un efecto contrario al esperado, ya que puede entorpecer el crecimiento del langostino.

Según Navas U. Juan (2018), las cualidades de un modelo son las siguientes:

- Coherencia; que guarde una relación lógica con la data pasada y permita prever el comportamiento de observaciones futuras.
- Generalización; que permita aplicarse la extrapolación de las relaciones determinadas por el modelo a otros sistemas que comparta las características mínimas del sistema en el que se generó el sistema, de acuerdo a los límites previamente establecidos.
- Robustez; es decir, que tenga la capacidad de respuesta ante los cambios de las variables de entrada.
- Flexible; porque debe permitir cambios y tener la capacidad de adaptabilidad de nuevos escenarios en el sistema.

3.1. Importancia del modelamiento

Un modelo, según el profesor Sixto Ríos (1995), es todo objeto, concepto o conjunto de relaciones que tiene como finalidad representar y estudiar de manera más simple una parte de la realidad.

Por lo tanto, se puede decir que un modelo es una construcción artificial que permite expresar una realidad concreta en términos matemáticos, que pretende reunir solo aquellas características que son importantes para el objetivo que motiva la construcción de dicho modelo.

El interés en el modelamiento de diferentes sistemas o fenómenos han ido en aumento durante los últimos años a causa del crecimiento exponencial de datos que se generan, el avance en los ordenadores y software matemáticos, la necesidad de competitividad y de predicción, tanto en campos como medicina, agronomía, cadena de suministros, etc. Su importancia se justifica a través de los aportes o resultados que se pueden obtener a partir de este, como son la comprensión y estudio de complejos sistemas, identificación de variables críticas del sistema para controlar, predicción del comportamiento bajo ciertas condiciones del sistema, etc.

Su importancia básicamente se concentra en la simplificación de una realidad en un modelo para entenderlo, debido a que los problemas reales son complicados, ya que los elementos que intervienen son numerosos además que las relaciones entre dichos elementos no son evidentes ni sencillos de identificar. Para la construcción de un modelo es necesario determinar los componentes, es decir, descartar aquellos elementos que no sean importantes, establecer las relaciones entre estos elementos, y finalmente construir la expresión gráfica o matemática del modelo que permitirá representar a la realidad en estudio. (Narro Ramirez, 1996)

En consecuencia, ante la necesidad de entender el comportamiento del fenómeno y representarlo de una manera más sencilla y sistemática, se presenta como una buena opción el modelamiento.

3.2. Técnicas de modelamiento

Las técnicas de modelamiento son diferentes tratamientos que se le da a los datos para obtener información que se utiliza para identificar tendencias y patrones.

Existen diferentes técnicas estadísticas, técnicas de aprendizaje automático de datos (machine learning), técnicas de minería de datos, etc. Los métodos de *machine learning* y minería de datos están basados en técnicas estadísticas, pero no pueden confundirse con ellas. Todas estas técnicas se relacionan, pero la diferencia está en el enfoque. Mientras que las técnicas de *machine learning* se enfoca en la predicción, las técnicas estadísticas en inferencias. Sin embargo, todas tienen el mismo propósito: aprender de los datos.

Es importante notar que, tras identificar la relación entre las variables, a través de un modelo, surge la necesidad de realizar las suposiciones correspondientes de manera que guarden relación y coherencia con la realidad, es decir, que expresen una lógica razonable.

Las aproximaciones que se usan para modelar suelen clasificarse en, estimaciones paramétricas y las no paramétricas. Las paramétricas se caracterizan por asumir la forma de la función y luego entrenar con la data a fin de definir un modelo sencillo. Tiene la desventaja de ser poco preciso para los ajustes, por ello para mitigar este problema se buscan funciones más

flexibles, aunque esto aumente el riesgo de caer en el denominado *overfitting*, es decir, que la forma de la función se distorsione como resultado de seguir a los errores o ruidos.

Los modelos obtenidos por métodos no paramétricos se caracterizan por no asumir ninguna función ya que se busca y construye la función según los datos; este método tiene la ventaja de encontrar la forma de la función estimada en un rango extenso de formas. Sin embargo, cuenta con la desventaja de no reducir el número de parámetros de la función y requiere una mayor cantidad de observaciones. (Gareth, Witten, Hastie, & Tibshirani, 2017)

En general, las diferentes técnicas del modelamiento de datos se pueden agrupar en técnicas de regresión y técnicas de aprendizaje computacional, esta clasificación se explicará líneas abajo. (Espino Timón, 2017)

3.2.1. Técnicas de regresión

Surge como respuesta ante la necesidad de expresar, de forma matemática, la relación entre las variables independiente, que son aquellas cuyo valor no depende de ninguna otra variable, y las variables dependientes, que son la consecuencia del efecto conjunto de cada una de las variables independientes. Existe una gran variedad de modelos aplicables, dependiendo de la situación en cuestión. A continuación, se describirá muy brevemente las más utilizadas.

La regresión lineal, es un arreglo matemático que permite establecer una ecuación para predecir el valor de la variable respuesta a partir de un análisis de las relaciones lineales entre las variables. Las variables de entrada son aquellas que contribuyen de manera significativa al cálculo de la variable respuesta, estas a su vez, son acompañadas por sus coeficientes que son parámetros calculados que se ajustan para obtener los resultados con el menor error posible. (Espino Timón, 2017)

La regresión no lineal, es una ecuación o conjunto de ecuaciones que busca modelar fenómenos que no pueden ser explicados de forma lineal, por lo tanto, permite modelar comportamientos más complejos debido al uso de expresiones matemáticas más complejas. Sin embargo, al igual que la regresión lineal cuenta con los mismos supuestos para los errores. (Rivas M., López, & Velasco, 2004)

Análisis de supervivencia, es el análisis del tiempo o duración para que suceda un evento, suele ser utilizada para el campo de la medicina, para evaluar mortalidades, para describir el tiempo de vida de dispositivos, herramientas, etc. Sin embargo, por la naturaleza de la duración, siempre positivo, no permite aceptar valores negativos, así que no es posible la suposición de la normalidad para este tipo de modelado. (Espino Timón, 2017)

Curvas de regresión adaptativa multivariable, es una técnica que permite modelar de manera flexible por medio de un conjunto de regresiones por secciones, así que se definen en la intersección de los modelos de regresión lo denominados nudos. (Espino Timón, 2017)

En este estudio se aplicará la regresión lineal y no lineal, los cuales resultaron más apropiados para modelar el fenómeno de interés.

3.2.2. Técnicas de aprendizaje computacional

El aprendizaje computacional es el área central de interés de la inteligencia artificial, que dirige sus esfuerzos al estudio y desarrollo de algoritmos que puedan mejorar su desempeño a medida que aumenta su experiencia, a través de la creación de modelos matemáticos sin la intervención de alguna asistencia humana, ya que el objetivo es conseguir un aprendizaje autónomo y automático de manera que emule el proceso cognitivo humano. Ha tomado mayor relevancia durante los últimos años debido a la disponibilidad de una gran cantidad de datos, maquinas con mayor capacidad de almacenamiento y procesamiento, y la facilidad para usar estas herramientas. (Pinedo Cortés, 2017)

A continuación, se describen brevemente algunos de los métodos más utilizados de *machine learning*. Redes neuronales, permite representar comportamientos complejos mediante el aprendizaje de la relación entre las variables de entradas y la de salida a partir de la experiencia que se obtiene a través de un entrenamiento, en el cual estas redes se comportan como el cerebro humano para procesar la información, mientras que, en paralelo se aprende y generaliza (Serna M., Serna, & Acevedo, 2017).

Máquinas de vectores de soporte, es una técnica de clasificación que se usa para detectar y explotar patrones de comportamientos complejos. Ha llamado mucho la atención debido a su desempeño para resolver problemas de clasificación, basándose en la reducción del riesgo estructural (Betancourt, 2005).

Naïve Bayes, es una técnica basada en la regla de probabilidad condicional de Bayes en conjunto con la hipótesis de independencia condicional de las variables predictoras, cuya importancia y reconocimiento se basa en su robustez para solucionar problemas de clasificación supervisada. (Larrañaga, Inza, & Moujahid, 2000)

k-vecinos más cercanos, es una técnica estadística orientada a la clasificación fundamentada. clasifica a los nuevos casos en la clase más frecuente a la que pertenecen sus k-vecinos más cercanos. (Moujahid, Inza, & Larrañaga, 2000)

3.3. Debate y selección de la técnica a utilizar

La selección de la técnica a utilizar deberá tener en su óptica el objetivo principal del estudio a realizarse. La inferencia, que se pretende obtener a partir del modelamiento permitirá entender la relación y la manera en cómo afectan cada una de las variables entrada a la variable de salida, es decir, establecer la medida de cambio de la variable respuesta según la variación de las variables de ingreso.

Los modelos obtenidos mediante la aplicación de técnicas de aprendizaje computarizado se caracterizan por adecuarse a los datos de entrenamiento, es por ello que tiende a generar relaciones con el fin de obtener el valor correspondiente a la respuesta, es decir, que se establecen estas relaciones entre variables como resultado de una serie de operaciones matemáticas que se realizan el interior de la denominada caja negra o "*black box*" porque desconocemos la razón del origen de estas relaciones y esto complica el segundo propósito del modelamiento, que es la interpretación.

Mientras que las técnicas por regresión permite obtener a partir del uso de herramientas estadísticas la construcción de modelos que pueden predecir los posibles resultado pero no con tanta precisión que la del aprendizaje computacional, pero a diferencia de esta, el modelamiento por regresión nos permite responder ciertas preguntas de interés para la inferencia del modelo, como, ¿cuáles de todo los predictores se encuentran realmente asociados a la variable respuesta?, ¿cuál es la relación entre cada uno de los predictores y la variable respuesta?, y ¿si las relaciones que existen entre las variables son lineales o mantienen relaciones más complejas?

Dependiendo de la importancia que se le atribuya a la predicción y la inferencia del modelo, se seleccionará la técnica a usar para el modelamiento. Entonces, el método que se elija debe tener en cuenta la compensación que existe entre la precisión del modelo y la interpretabilidad de este, aunque aparentemente sea más apropiado utilizar métodos no paramétricos debido a la flexibilidad de sus estimaciones, puede que las aproximaciones paramétricas sean más adecuadas y por las cuales se decida trabajar con estas.

La sencillez e inflexibilidad de las aproximaciones paramétricas suelen permitir obtener modelos más fáciles de entender, en cambio las aproximaciones no paramétricas debido a su flexibilidad inciden en complejos estimadores para la función deseada, esto complica la identificación de la relación de cada una de las variables individuales con la variable respuesta. No existe el método estadístico perfecto que pueda tener la mejor precisión en cuanto a la predicción, interpretabilidad, flexibilidad, supervisión de las variables y observaciones, es por este motivo que, según sea el particular objetivo del modelado se debe elegir aquel método de aproximación más conveniente. (Gareth, Witten, Hastie, & Tibshirani, 2017)

Esta compensación entre interpretabilidad y precisión se observa también en el error cuadrático medio, MSE por sus siglas en inglés, es el indicador que permite reconocer que tanto se acercan las estimaciones a los valores reales. Esto permite identificar qué modelo es más preciso para las predicciones, pero también es muy probable que aquel pueda caer en el error del *overfitting*, perder el de vista el comportamiento medio por seguir de cerca a las observaciones.

Este indicador estadístico, se puede desglosar en tres términos, el primero es error del modelo estimado porque hay error generado por el modelo estimado que se ha elegido, el segundo es el sesgo que existe entre los estimadores y los parámetros reales porque dependiendo de la calidad de los datos se obtendrán estos estimadores, y tercero es la variabilidad propia del error irreducible. Esto hace suponer que mientras más flexible sea la estimación menor será su sesgo, ya que la estimación de los parámetros sigue más de cerca a las observaciones, pero la varianza terminará por aumentar. (Gareth, Witten, Hastie, & Tibshirani, 2017)

Ante el mayor interés por la inferencia del modelo, para este estudio se ha creído conveniente realizar un modelo de dos fases. En la primera fase se pretende modelar el comportamiento medio no lineal del crecimiento del langostino a través de un modelo no-lineal. En la segunda fase se usará un modelo de regresión lineal múltiple con el objetivo de descubrir o identificar aquellos factores que pueden causar la desviación con respecto a ese comportamiento medio modelado en la primera fase.

3.4. Modelo de dos fases

Frente al complejo fenómeno del crecimiento del langostino este estudio propone la estimación paramétrica de un modelo de dos fases para identificar y comprender aquellos factores que intervienen en él.

Este modelamiento se compone por una primera fase, en la cual se busca modelar el comportamiento medio del crecimiento del langostino en el tiempo sin la intervención de ningún otro factor que afecte este desarrollo, y una segunda fase en la cual a partir de los residuos, resultado de la anterior fase, se pretende construir un modelo de regresión lineal para de esta manera reconocer aquellas las fuentes de variabilidad que explican las desviaciones que tenga cada piscina con respecto a ese comportamiento medio.

La Figura 17 muestra la gráfica entre el peso y el número de semanas, según sus respectivas piscina, vemos que el langostino crece durante las primeras 8 semanas a un ritmo medio más lento que en las semanas comprendidas entre la 8 y la 15, para luego pasar a un periodo de receso en el cual su crecimiento se ralentiza porque alcanza ya un tamaño máximo, esto resulta en

gráficas cuya razón de crecimiento son diferentes según el número de semana. Es por este comportamiento que su gráfica tiene una forma de "S" alargada.

Sin embargo, esta forma presenta los datos con un comportamiento más desordenado en cuanto a la progresión del crecimiento. Para reducir esta diferencia en cuanto a la escala del desarrollo del crecimiento entre las piscinas, se aplica una transformación con el logaritmo natural. Como se muestra en la Figura 18 los datos se uniformizan, y las razones de crecimiento tienen una progresión más uniforme, además que permite visualizar de una mejor manera el comportamiento asintótico del peso máximo. Además, es conveniente utilizar esta transformación debido a que la construcción de modelos con logaritmo permite la interpretación del porcentaje de cambio, esto se explicará más adelante en el presente capítulo de modelamiento.

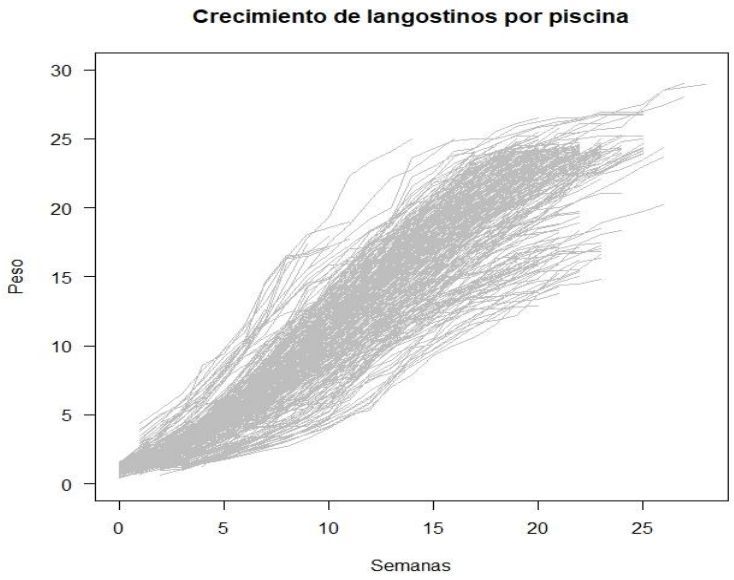


Figura 17. Gráfica del peso vs el número de semana por semana. Fuente: elaboración propia, con información recuperada de la empresa ECOSAC.

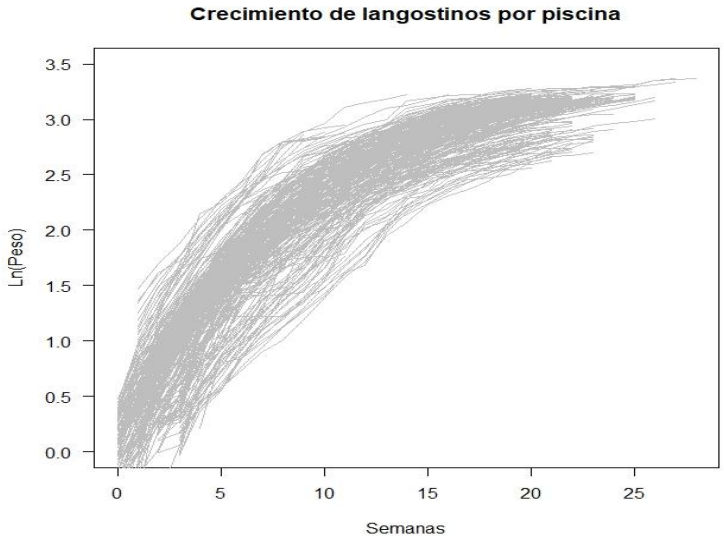


Figura 18. Gráfica del Ln peso vs el número de semana por piscina. Fuente: elaboración propia, con información recuperada de la empresa ECOSAC.

El modelo de dos fases se pude definir de la siguiente forma general:

$$log\,weight_{ij} = f(x_{ij}, \boldsymbol{\theta}) + \epsilon_{ij} \quad (1)$$

$$\epsilon_{ij} = g(\boldsymbol{Z}_{ij}, \boldsymbol{\beta}) + \zeta_{ij} \qquad i = 1, \dots t\ ,\ j = 1, \dots m \quad (2)$$

donde:

- $f(x_{ij}, \boldsymbol{\theta})$ es una función no lineal con el vector de parámetros $\boldsymbol{\theta}$, x_{ij} representa el tiempo en semanas en donde *i* representa el número de semanas en la piscina *j*.
- $g(\boldsymbol{Z}_{ij}, \boldsymbol{\beta})$ es una función lineal con el vector de parámetros $\boldsymbol{\beta}$, y Z_{ij} representa todas las variables que pueden influenciar el crecimiento del langostino.

$$\zeta_{ij} \sim N(0, \sigma^2)$$

La primera fase tiene como finalidad modelar el efecto que solo se le atribuye al crecimiento normal del langostino con respecto al tiempo, suponiendo que el alimento no es un limitante, se utilizará un modelo no lineal, y la segunda utilizará un modelo de regresión lineal que pretenda explicar el efecto de las demás variables que interactúan con el crecimiento del langostino haciendo de los residuos de la primera fase.

3.4.1. Modelo no-lineal

En primera instancia se evaluará el comportamiento del crecimiento medio del langostino en función del tiempo, es decir se considerará, un modelo que pretenda estimar el peso ajustando un modelo que se adecue, lo mejor posible, a los datos que se disponen.

El tiempo se expresará en número de semanas que tenga el langostino en las piscinas de engorde y el peso se expresa en gramos, variable de mayor interés porque servirá como indicador del crecimiento del langostino.

En la literatura se han estudiado diferentes modelos no lineales para representar el crecimiento de los animales, los cuales parten de la suposición que el crecimiento es proporcional a su peso según su edad, estas funciones suelen ser, por ejemplo, *Schumacher, Logistic y Gompertz y Von Bertalanffy,* de este último profundizará más líneas a bajo.

Schumacher-model es un modelo de parámetros de forma y escalas variables que, es un modelo del tipo logarítmico y cuya tendencia no necesariamente es exponencial, posee una buena capacidad predictiva, y se caracteriza por tener un cambio más ajustado en la pendiente en las curvas del tipo sigmoidal. Inicialmente fue desarrollado para relacionar el volumen con la edad bajo el supuesto que el crecimiento poblacional se relaciona de forma inversa con la edad. (López Hernández & Valles Gándara, 2009)

$$F(t) = \beta_0 * e^{-\beta_1 * (\frac{1}{t})} \tag{3}$$

Donde:

- β_0 y β_1 parámetros estimables.
- t edad o tiempo de vida.

Logistic-model es una función muy utilizada para modelar fenómenos de crecimiento que presenten una forma de curva sigmoidal, que se caracterizan por presentar 4 regiones. La primera región es de un crecimiento lento, seguido por la segunda zona que tiene un crecimiento mucho más acelerado, la tercera zona tiene un crecimiento lento con tendencia a estabilización, y finalmente la cuarta zona que es la estabilización de la curva debido a alcanzar su valor asintótico. En consecuencia, a estas características el modelo resulta ser muy versátil, lo suficiente, como para ser utilizado para modelar el crecimiento poblacional, número de bacterias, tiempo de respuesta a medicamentos, etc. Se diferencia al anterior por la presencia de una constante "k" que representa el límite o el punto de saturación del sistema. (Ulloa Ibarra & Rodríguez Carrillo, 2010)

$$P(t) = \frac{kP_0e^{rt}}{k + P_0(e^{rt} - 1)} \tag{4}$$

Donde:

- P_0 es la población inical

- r es el valor que se varía hasta que la gráfica se ajuste a los datos

Gompertz-model es, según (Casas, Rodríguez, & Afanador Téllez, 2010), un modelo no linear utilizado principalmente para modelar fenómeno biológicos y económicos que se encuentran relacionados con el crecimiento, el cual presume que este crecimiento en los individuos se realiza en un progresión constante hasta llegar a un valor máximo de crecimiento para luego iniciar su decrecimiento asintótico hasta alcanzar su tamaño máximo. Modela comportamientos que presenten una forma de curva sigmoidal, con un punto de inflexión, que representa la tasa máxima de crecimiento, y una asíntota que representa el peso o tamaño máximo. Se diferencia al modelo Schumacher por su grado de flexibilidad como la función *Logistic*, pero a diferencia de este último en su punto de inflexión debido a que no se encuentra a la mitad de la curva sino antes, es decir, que modela curvas sigmoidales no simétricas. (Winsor, 1932)

$$weight_i = f(t) = \alpha e^{-e^{-\kappa(t-\gamma)}} \tag{5}$$

Donde:

- α es el peso final o peso máximo (asíntota)
- κ gobierna el ratio de crecimiento
- γ punto de inflexión

Esta función, *Gompertz-model*, se ha utilizado para modelar el crecimiento o desarrollo de animales[4], plantas, etc. (Hassan Darmani kuhi, 2010), en función al tiempo, evaluando tres parámetros el ratio de crecimiento, tamaño máximo y punto de inflexión. Debido a que este modelo representa de la mejor manera el crecimiento proporcional al peso, ya que inicialmente tiene un crecimiento exponencial y luego un crecimiento asintótico.

Von Bertalanffy-model es un modelo de comportamiento exponencial que es frecuentemente utilizado para modelar el crecimiento en la biología marina, con la finalidad de relacionar la edad con la altura de las especies marinas. Este modelo se caracteriza por reconocer los crecimientos que no mantienen un ritmo constante a lo largo del ciclo de vida y por considerar una tasa de crecimiento acelerada en la etapa inicial, que a medida se acerca a la etapa adulta irá reduciéndose hasta llegar a una etapa de crecimiento nulo. La función matemática define la altura o peso máximo cuando el tiempo tiende al infinito, así también a la razón de crecimiento como proporcional a la altura o peso, además que gracias a sus tres parámetros consigue una relativa

[4] Wright (1926), Medawar (1940), y Seber and Wild (2003)

flexibilidad mediante. Busca conseguir la curva media mediante la determinación de estos parámetros por el método de mínimos cuadrados. (Científica, 2011)

$$L(t) = L_{\infty}(1 - e^{-r(t-t_0)}) \quad (6)$$

Donde:

- L_{∞} es la longitud cuando el tiempo tiende al infinito.
- r es la proporción de la tasa de crecimiento y la longitud que le falta por crecer.

A continuación, se discute sobre la elección del modelo no lineal para usar en la primera fase del modelado.

De los modelos que se mencionaron anteriormente, los que más se utilizan para modelar el crecimiento de especies marinas son la *Logistic-model, Gompertz-model, Von Bertalanffy-model*. Estos modelos se utilizan para representar a curvas sigmoidales, sus diferencias radican en la estimación de esta curva y la flexibilidad que cada modelo tiene para hacerlo, *Gompertz-model* tiene una mayor flexibilidad que el modelo de *Logistic-model* a causa de su punto de inflexión, mientras que el *Von Bertalanff-model* es apropiada y en consecuencia más utilizada para el modelamiento del comportamiento animal por ser una función derivada en base a propiedades biológicas de crecimiento (Goodall & Sprevak, 1984).

Según una evaluación de modelos hecho por Xijun Tian & Hochman (1993), los dos que se comportan de mejor manera para modelar el crecimiento del langostino son *Gompertz-model* y *Von Bertalanffy-model* según una evaluación de bondad de ajuste, mínimos residuos, peso asintótico y punto de inflexión, donde se identificó que el *Logistic-model* sigue de forma más precisa, por una sección de la curva, a las observaciones reales, esto debido a su punto de inflexión. Sin embargo, su valor de peso asintótico se encuentra por debajo del peso estándar máximo al que llegan los langostinos, entonces, se le descarta como una buena función para modelar el crecimiento de los langostinos.

Según Araneda, Hernández, Gasca-Leyva, & Vela (2013), la función *Von Bertalanffy* representa de la mejor manera la evolución homogénea del peso de los langostinos, debido tener un valor de *theil´s U* menor al crítico de 0.2 y un bajo valor de error cuadrático medio, así también, porque permite modelar la función completa. Es decir, permite alcanzar el valor de peso medio máximo, valor asintótico o estable de un langostino maduro, que no permite la función de *Gompertz* porque tiende a determinar un valor asintótico menor al real. Además, que la función de *Von Bertalanffy* cuenta con parámetros que son mucho más fáciles de interpretar de forma física, es decir, que contribuye a entender mejor la evolución de obtención de peso de los langostinos.

Quevedo, Vegas, Loda, Cedino, & Vining, Within batch non-linear profile monitoring applied to shrimp farming (2020) se centra en el modelo *Von Bertalanffy*, el cual indica que si no tiene ninguna restricción en cuanto a la alimentación se puede considerar el crecimiento proporcional al peso, además que se adecua mejor a modelos en los cuales el crecimiento disminuye a medida que el tiempo transcurre.

Como se mencionó anteriormente se realizó la transformación de la variable peso, con la operación de logaritmo natural, para obtener una expresión en función del peso cuya forma sea más sencilla de modelar. La función de *Von Bertalanffy-model* ha sido utilizada para modelar el crecimiento del langostino en varios estudios (Wright (1926), Medawar (1940), Tian et al. (1993), Seber and Wild (2003), Yu et al. (2006) y Araneda et al. (2013)), debido a ser una función cuyos parámetros son mucho más sencillos de interpretar.

Por todo lo anteriormente comentado, en este estudio se usó la función de *Von Bertalanffy*, de la siguiente manera:

$$\log weight_{ij} = f(x_{ij}, \boldsymbol{\theta}) = \theta_1 - \theta_2 e^{-\theta_3 x_i} + \epsilon_{ij}\,, \qquad i = 1, \dots t\,, j = 1, \dots m \tag{7}$$

Donde:

- x_{ij} es la semana número i para piscina j.
- θ_1 representa el logaritmo del peso cuando t → ∞, es decir, el peso máximo.
- θ_3 es el ratio de crecimiento.
- $\theta_1 - \theta_2$ representa el logaritmo del peso en el tiempo $t = 0$.
- ϵ_{ij} es el error.

Para estimar este modelo no-lineal se puede usar cualquier software estadístico. El presente estudio ha utilizado Minitab y el software R. Para estimar cualquier modelo no-lineal se debe ingresar los valores iniciales de los parámetros del modelo. Los valores iniciales de los parámetros son $\theta_1 = 3.8$, indica que el logaritmo natural del peso máximo que alcanza un langostino maduro que puede ser aproximadamente 48 g, $\theta_2 = 3.3$, y $\theta_3 = 0.1$. Estos sirven cómo parámetros de referencia para, que con un algoritmo de iteración que realiza el software, usando la campaña 15, se estime el siguiente modelo no-lineal:

$$\ln(peso) = 3.4201 - 3.34968 * \exp\,(-0.110996 * N°Semana) \tag{8}$$

Los parámetros estimados de la curva que mejor se ajusta a la data coinciden en ambos software, esto refuerza que el modelo estimado sí representa a la curva media de entre todas las

curvas de crecimiento por piscina, estos parámetros son $\theta_1 = 3.4201065$, lo cual indica que el peso máximo del langostino maduro es de aproximadamente 30.57247 g, $\theta_2 = 3.3496805$, y $\theta_3 = 0.1109952$, es decir que tiene un ratio de crecimiento medio de 1.11739 g por semana. Es importante mencionar que se ha utilizado la data a partir de la semana 1 ya que la mayoría de las piscinas no tienen registro de la semana 0.

El modelo representa el comportamiento asintótico del crecimiento de los langostinos tal como se muestra en la Figura 19, ajustándose bastante bien a los datos. Los residuos se pueden observar en la Figura 20, son las diferencias entre el valor observado y el valor obtenido por la función no lineal, estos residuos son importantes para la siguiente parte del modelamiento.

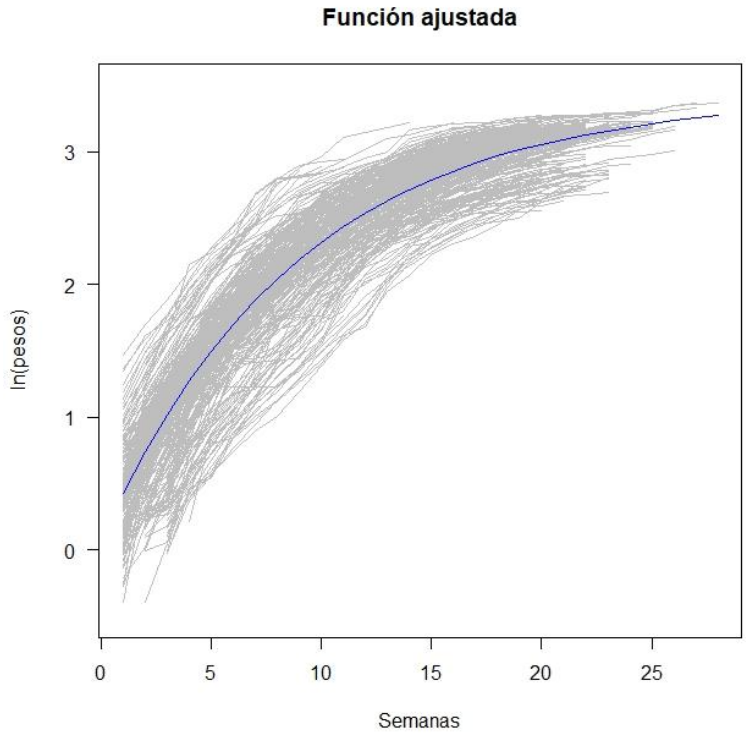


Figura 19.Gráfica de la curva ajusta a los datos en RStudio.
Fuente: elaboración propia, con información recuperada de la empresa ECOSAC.

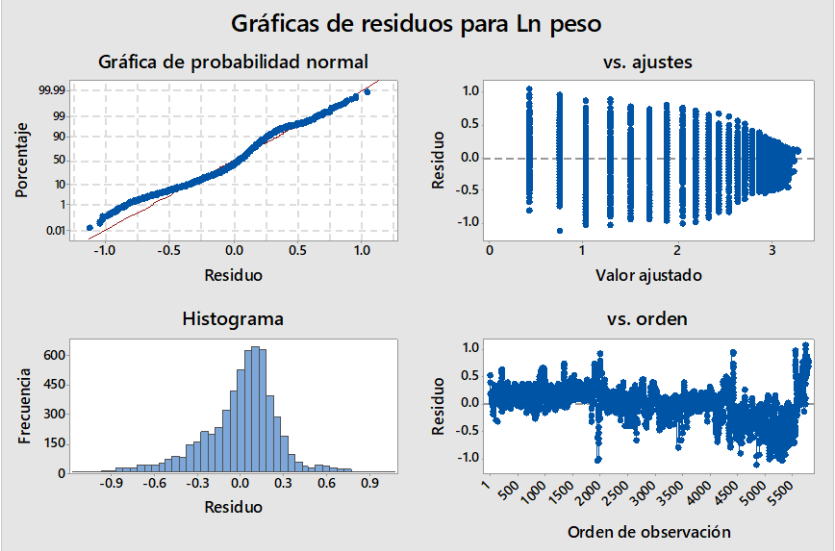


Figura 20. Gráficas de resumen de los residuos.
Fuente: elaboración propia, con información recuperada de la empresa ECOSAC.

Los residuos que se obtienen luego de ajustarse a la regresión no lineal (se pueden interpretar como las desviaciones con respecto al comportamiento medio estimado. En la siguiente fase, se quiere pretende estudiar, la influencia que pueden tener aquellos factores sobre estas desviaciones.

3.4.2. Modelo de regresión lineal múltiple

La segunda parte de este modelo se basa en entender a los residuos generados por el modelo anterior. En este caso asumiremos que esta variable tendrá un comportamiento lineal, es decir, que las desviaciones con respecto a ese comportamiento medio se pueden estimar mediante una combinación lineal de aquellos factores que pueden estar afectando dicho comportamiento. Estas posibles variables críticas fueron descritas en el capítulo de Limpieza de datos, Capítulo 2.

Esta segunda parte del modelo de dos fases se puede expresar, como se mostró en la ecuación (2), de la siguiente manera:

$$\epsilon_{ij} = g(\boldsymbol{Z}_{ij}, \boldsymbol{\beta}) + \zeta_{ij}$$

$$\epsilon_{ij} = \beta_0 + \beta_1 z_{1ij} + \beta_2 z_{2ij} + \cdots + \zeta_{ij} \quad (9)$$

Para la semana $i = 1, \ldots t$, y para la piscina $j = 1, \ldots m$ $\zeta_{ij} \overset{iid}{\sim} N(0, \sigma^2)$

Donde:

- $\boldsymbol{Z}_{ij}$ son las variables independientes que se presume pueda afectar al crecimiento.
- $\boldsymbol{\beta}$ son los coeficientes de la regresión lineal multivariable.
- ζ_{ij} son los errores que corresponden a cada respuesta estimada.

Variables

Se evaluarán los parámetros anteriormente descritos en el Capítulo 2, entre los cuales destacan los siguientes:

- Mes de siembra
- Densidad inicial
- Porcentaje de Nicovita, Gisis y Purina
- Alimento semanal
- Relación de alimento vs total de animales
- Factor de conversión semanal
- Factor de conversión acumulado
- Kw por animal
- Temperatura
- Oxígeno
- pH
- Salinidad
- Alcalinidad CaCO3.
- Cianofitas, Clorofitas y Diatomeas

Además, se evaluarán las siguientes variables categóricas:

Variable categórica del mes**:** según se muestra en las Figura 21 y Figura 22 la temperatura tiene un comportamiento oscilante a medida que el mes del año avanza, entonces, esta variable categórica permite representar el efecto de las condiciones, tratamiento y decisiones que se han tomado según su mes y si este afectó a las variables anteriormente mencionadas.

Con esta variable se pretende entregar información que no se pueda expresar por alguna variable o de la cual no se tenga información, y que se aísle este efecto distinto que podría haber entre los grupos correspondientes según su mes de siembra.

Se dispusieron diferentes formas de agrupar los meses los cuales se definieron según el comportamiento de la temperatura, como se muestran en la Figura 23 y Figura 24 la temperatura, tanto mínima como máxima, es por este motivo que se ha decidido agrupar de la siguiente manera y teniendo sus correspondientes consideraciones:

Tabla 2. Agrupación de los meses de siembra según sus temperaturas

Categórico mes	0	1	2
Categórico-mesC	Enero, febrero, marzo y abril	Mayo, junio, julio y agosto	Septiembre, octubre, noviembre y diciembre

Fuente: elaboración propia, con información recuperada de la empresa ECOSAC

Estas agrupaciones de los meses de siembra, que se muestran en la Tabla 2, se hacen teniendo en consideración las temperaturas a las que se encuentran las piscinas a lo largo del ciclo de vida del langostino, entonces el criterio de la categórica son los valores de temperatura según su mes de siembra.

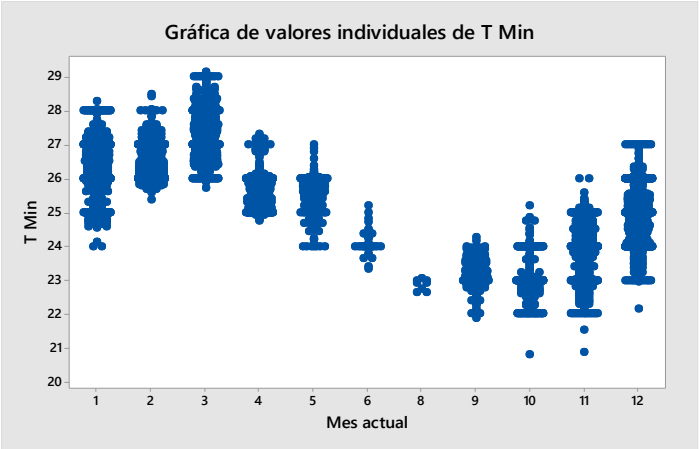


Figura 21. Gráfico de valores individuales de temperatura mínimo. Fuente: elaboración propia, con información recuperada de la empresa ECOSAC.

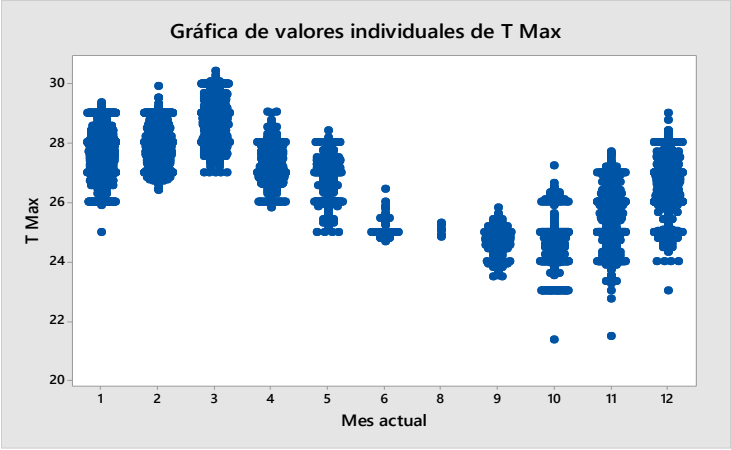


Figura 22. Gráfico de valores individuales de temperatura máximo. Fuente: elaboración propia, con información recuperada de la empresa ECOSAC.

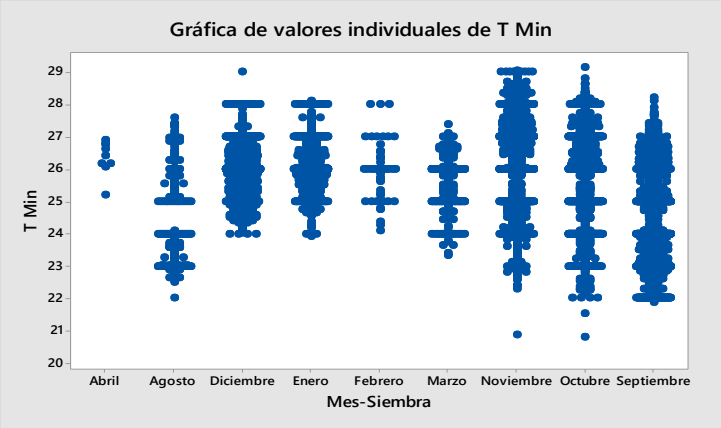


Figura 23. Gráfico de valores individuales de Temperatura mínima vs mes de siembra.
Fuente: elaboración propia, con información recuperada de la empresa ECOSAC.

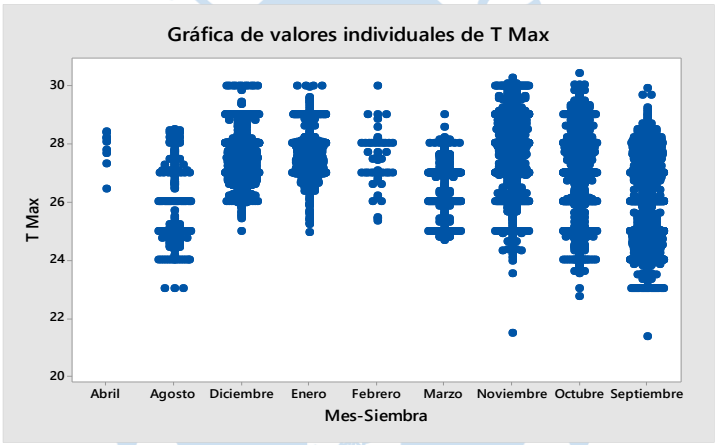


Figura 24. Gráfico de valores individuales de Temperatura mínima vs mes de siembra.
Fuente: elaboración propia, con información recuperada de la empresa ECOSAC.

Variable categórica del tipo de piscina: Es un tipo de variable que al igual que al anterior busca aislar un posible trato distinto entre las piscinas normales, que tomarán el valor de "0", y las piscinas que tienen por nombre reservorio, que toman el valor de "1", debido a que se sospecha que tienen una forma tratamientos distintos que pueden influir en el crecimiento del langostino.

Tras la evaluación de las variables continuas y las variables categóricas del mes y del tipo de piscina, se identificaron aquellas significativas para el modelo, y con las cuales se conseguía un mayor valor de R cuadrado, menor MSE y el bajo VIF (factor de inflación de la varianza) que expresa la magnitud de la multicolinealidad. Los valores de los coeficientes del modelo de regresión son pequeños debido a la magnitud de los errores, ya que toman valores entre -1 y 1. A continuación, se presenta el modelo de regresión resultante:

Después de este arduo análisis de posibles modelos para evaluar qué incluir y cuáles no, se seleccionó el siguiente modelo de regresión múltiple cuya especificación se muestra en Figura 25:

$$\hat{\epsilon} = -2.0343084 + 0.0412679 * \%\ Nicovita + 0.1069514 * pH\ Min - 0.0172229 * O2\ Max + 0.0405461 * T\ Max - 0.0172375 * Cianofitas(100K) + 0.0014434 * Alcalinidad\ CaCO3 - 0.1008640 * Factor\ de\ conversión + 0.1894017 * Car\_piscina1 + 0.1165539 * Cat\_mes1 + 0.2192120 * Cat\_mes2 \quad (10)$$

Categórica tipo de piscina	Categorico-mesC			
0	0	Resid	=	-2.034 + 0.0413 % Nicovita + 0.1070 pH Min - 0.01722 O2 Max + 0.04055 T Max - 0.01724 Cianofitas (100K) + 0.001443 Alcalinidad CaCO3 - 0.1009 Factor de conversión
0	1	Resid	=	-1.918 + 0.0413 % Nicovita + 0.1070 pH Min - 0.01722 O2 Max + 0.04055 T Max - 0.01724 Cianofitas (100K) + 0.001443 Alcalinidad CaCO3 - 0.1009 Factor de conversión
0	2	Resid	=	-1.815 + 0.0413 % Nicovita + 0.1070 pH Min - 0.01722 O2 Max + 0.04055 T Max - 0.01724 Cianofitas (100K) + 0.001443 Alcalinidad CaCO3 - 0.1009 Factor de conversión
1	0	Resid	=	-1.845 + 0.0413 % Nicovita + 0.1070 pH Min - 0.01722 O2 Max + 0.04055 T Max - 0.01724 Cianofitas (100K) + 0.001443 Alcalinidad CaCO3 - 0.1009 Factor de conversión
1	1	Resid	=	-1.728 + 0.0413 % Nicovita + 0.1070 pH Min - 0.01722 O2 Max + 0.04055 T Max - 0.01724 Cianofitas (100K) + 0.001443 Alcalinidad CaCO3 - 0.1009 Factor de conversión
1	2	Resid	=	-1.626 + 0.0413 % Nicovita + 0.1070 pH Min - 0.01722 O2 Max + 0.04055 T Max - 0.01724 Cianofitas (100K) + 0.001443 Alcalinidad CaCO3 - 0.1009 Factor de conversión

Figura 25. Ecuación de regresión del modelo lineal múltiple para combinación de valores de las variables categóricas.
Fuente: elaboración propia, con información recuperada de la empresa ECOSAC.

El modelo cumple que todas sus variables son significativas con un nivel de significancia de 0.05 tal como se muestra en la Figura 26. Cabe recalcar que, según el análisis de coeficientes del modelo lineal en la Figura 27, los valores de los VIF se encuentran bastante bajos esto significa que existe una baja colinealidad entre los variables predictoras, es decir, que ninguno de los regresos (variables de ingreso) resultan de la combinación lineal de los restantes.

Análisis de Varianza

Fuente	GL	SC Ajust.	MC Ajust.	Valor F	Valor p
Regresión	10	26.3123	2.63123	87.00	0.000
% Nicovita	1	0.3304	0.33040	10.93	0.001
pH Min	1	2.8706	2.87057	94.92	0.000
O2 Max	1	0.3074	0.30744	10.17	0.001
T Max	1	2.2544	2.25437	74.54	0.000
Cianofitas (100K)	1	4.6707	4.67071	154.44	0.000
Alcalinidad CaCO3	1	1.9327	1.93271	63.91	0.000
Factor de conversión	1	0.6974	0.69738	23.06	0.000
Categórica tipo de piscina	1	1.7877	1.78769	59.11	0.000
Categorico-mesC	2	3.2445	1.62224	53.64	0.000
Error	896	27.0976	0.03024		
Total	906	53.4099			

Figura 26. Análisis de la varianza del modelo lineal
Fuente: elaboración propia, con información recuperada de la empresa ECOSAC.

Término	Coef	EE del coef.	Valor T	Valor p	FIV
Constante	-2.034	0.179	-11.36	0.000	
% Nicovita	0.0413	0.0125	3.31	0.001	1.14
pH Min	0.1070	0.0110	9.74	0.000	1.34
O2 Max	-0.01722	0.00540	-3.19	0.001	1.16
T Max	0.04055	0.00470	8.63	0.000	1.87
Cianofitas (100K)	-0.01724	0.00139	-12.43	0.000	1.17
Alcalinidad CaCO3	0.001443	0.000181	7.99	0.000	1.06
Factor de conversión	-0.1009	0.0210	-4.80	0.000	1.80
Categórica tipo de piscina					
1	0.1894	0.0246	7.69	0.000	1.05
Categorico-mesC					
1	0.1166	0.0389	2.99	0.003	1.50
2	0.2192	0.0220	9.96	0.000	1.61

Figura 27. Análisis de coeficientes del modelo lineal
Fuente: elaboración propia, con información recuperada de la empresa ECOSAC.

Por la teoría se sabe que los residuos de los modelos lineal siguen un comportamiento normal y centrado en cero, esto se pude comprobar en la Figura 28.

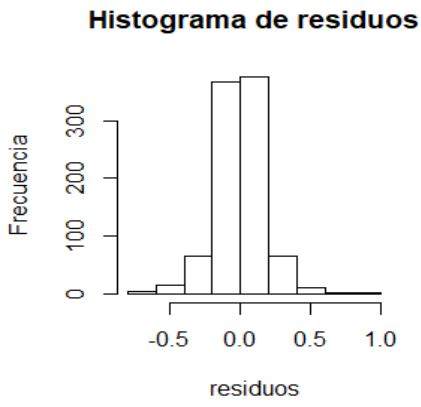


Figura 28. Histograma de residuos del modelo lineal
Fuente: elaboración propia.

Se realizó el mismo procedimiento en RStudio para obtener resultados más precisos de la regresión y respaldar el modelo hallado anteriormente. La regresión lineal múltiple se definió con los coeficientes que se muestran en la

Tabla 3, donde también se muestran los valores de sus respectivos VIF que verifican la baja colinealidad entre las variables y el p-valor que las reconoce como variables críticas.

Tabla 3. Tabla resumen de la regresión lineal múltiple

Variable	Coeficiente	Std. Error	VIF	p-val
(Intercept)	-2.0343084	0.1790514		< 2e-16
% Nicovita	0.0412679	0.1790514	1.142714	0.000986
pH mínimo	0.1069514	0.0124855	1.344161	< 2e-16
O2 máximo	-0.0172229	0.0109777	1.161527	0.001480
T máxima	0.0405461	0.0054018	1.867458	< 2e-16
Cianofitas	-0.0172375	0.0046962	1.167315	< 2e-16
Alcalinidad	0.0014434	0.0013871	1.064881	4.01e-15
Fact. Conver	-0.1008640	0.0001806	1.802742	1.84e-06
Cat. Piscina=1	0.1894017	0.0210042	1.054360	3.92e-14
Cat. mesC=1	0.1165539	0.0246349	1.290491	0.002824
Cat. mesC=2	0.2192120	0.0389217		< 2e-16

Fuente: elaboración propia. Estimación en software RStudio.

A partir de la data trabajada, este es el modelo más adecuado, así mismo este contempla las variables críticas para el crecimiento del langostino. Este modelo nos ayuda a entender el impacto de cada una de las variables que se encuentran incluidas.

Interpretación de los coeficientes

Como se mencionó anteriormente la transformación con el logaritmo, de la variable respuesta, ofrece muchas bondades útiles para la interpretación del modelo. Los coeficientes del modelo de regresión lineal se pueden interpretar como el porcentaje de cambio respecto al peso medio, por efecto de la transformación inicial. Por ello para el análisis del porcentaje de cambio se tratará al residuo como el logaritmo de una variable ficticia $\epsilon_{ij} \approx \log \epsilon_i$, para que se visualice de mejor manera la relación que tiene ϵ_{ij} con el peso medio del modelo no lineal de la primera fase, en la escala original de la variable respuesta $weight_{ij}$.

$$log\, weight_{ij} = f(x_{ij}, \boldsymbol{\theta}) + \epsilon_{ij} \rightarrow weight_{ij} = exp\left(f(x_{ij}, \boldsymbol{\theta})\right) * exp(\epsilon_{ij})$$

Se ha tomado como referencia el análisis hecho por D. Angrist & Pischke (2015, pág. 94), que busca evidenciar la interpretación del porcentaje de cambio, se parte con la ecuación lineal de la segunda fase (9):

$$log\ \epsilon_i = \beta_0 + \beta_1 z_{1ij} + \beta_2 z_{2ij} + \cdots + \zeta_{ij}$$

Si se asume por ejemplo que (z_{1ij}) es una variable binaria (que toma valores de ceros o unos). Para cualquier piscina "j" dejando a las demás constantes, y se evalúa $z_{10j} = 0$ se obtiene como resultado $log\ \epsilon_0$ y cuando $z_{11j} = 1$ se obtiene $log\ \epsilon_1$, se debe recordar que estas variables mantienen una naturaleza de logaritmo, por este motivo al evaluarse esto en la ecuación anterior lineal anterior se obtienen las siguientes expresiones:

$$log\ \epsilon_0 = \beta_0 + \beta_1(0) + \beta_2 z_{2ij} + \cdots + \zeta_{ij}$$

$$log\ \epsilon_1 = \beta_0 + \beta_1(1) + \beta_2 z_{2ij} + \cdots + \zeta_{ij}$$

Entonces, a partir de la resta de estas dos expresiones se puede llegar a la siguiente relación:

$$log\left(\frac{\epsilon_1}{\epsilon_0}\right) = \beta_1$$

$$log\left(\frac{\epsilon_1}{\epsilon_0}\right) = log\left(1 + \frac{\epsilon_1 - \epsilon_0}{\epsilon_0}\right) = log(1 + \Delta\%\epsilon_z) \approx \Delta\%\epsilon_z$$

La expresión que se obtiene tras el desarrollo es muy aproximada a la variación porcentual de los resultados potenciales inducidos por una variación en la variable z_{ij}. A partir de esto, se puede concluir que la pendiente de un modelo con logaritmo da un aproximado del cambio porcentual en la variable ϵ_i generado por un cambio en la variable predictora z_{ij}.

Ahora, esta conclusión se puede expandir a si se trata de una variable continua. La expresión anterior estaría afectada por la variación la variable predictora, ya no por la unidad como con las variables binarias, y la expresión quedaría de la siguiente manera: $log\left(\frac{\epsilon_1}{\epsilon_0}\right) = \beta_1 \Delta z$. Esto se puede resumir como la derivada parcial de la función lineal con respecto a la variable de cambio deseada $\frac{\partial y}{y} = \beta_1 \partial z$.

Si se trabaja esa relación identificada, ahora para variables continuas, podemos calcular el cambio porcentual generado por la variación en las variables, determinadas como significativas para el modelo lineal, se puede obtener la siguiente expresión:

$$log\left(\frac{\epsilon_1}{\epsilon_0}\right) = \beta_1 \Delta z \rightarrow \frac{\epsilon_1}{\epsilon_0} = exp\ (\beta_1 \Delta z)$$

$$\frac{\epsilon_1 - \epsilon_0}{\epsilon_{0j}} = exp\ (\beta_1 \Delta z) - 1$$

Al aplicar la función exponencial a los coeficientes multiplicado por la variación en la variable a evaluar, conseguimos que ya no se haga énfasis en el valor medio aritmético, como se está acostumbrado, sino que se enfatice en el valor medio geométrico, es decir, que la interpretación se enfoque en el ratio de cambio más que en evaluar de forma aritmética los coeficientes. En cuanto al término de intercepción representa el valor medio esperado de la variable respuesta ϵ_{ij} cuando el resto de las variables valen "0".

El coeficiente de la variable Nicovita es 0.0412679, es decir, tiene un efecto positivo frente al crecimiento medio esperado del langostino. Si esta variable toma el valor de "1", lo cual significa que todo el alimento está compuesto por Nicovita, se espera obtener un aumento de 0.0412679 en la curva del logaritmo del peso medio de la primera fase, lo cual equivale al ratio del valor medio geométrico de la variable respuesta $weight_{ij}$ que hay entre la variable que se evalúa y el resto de las variables.

Este ratio de cambio es $exp(\beta_1 \Delta z) - 1 = 0.042131$ expresado en porcentaje es 4.21% es decir que, en la escala original, la unidad en esta variable representa una variación del 4.21% respecto a la curva media del peso del langostino. En la Figura 29 se muestra la relación entre los residuos y la variable Nicovita.

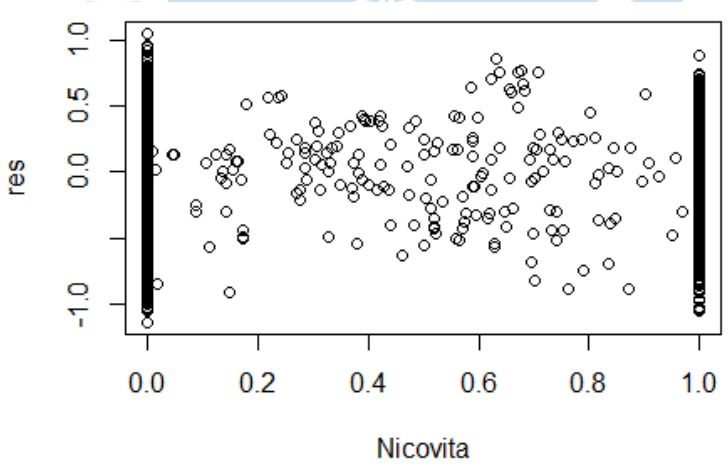


Figura 29. Gráfica de residuos vs Nicovita.
Fuente: elaboración propia. Estimación en software R

El pH mínimo semanal tiene una relación positiva con el crecimiento del langostino. Como vemos en la siguiente Figura 30, se observa una clara relación lineal. Según los resultados obtenidos, en promedio se espera un incremento del 0.1069514 (en logaritmo del peso) sobre el crecimiento medio del langostino si el pH del agua aumenta una unidad, según lo explicado anteriormente, este coeficiente representa un porcentaje de cambio relacionado al valor medio geométrico en la variable respuesta original.

Este ratio, según lo determinado anteriormente, es $exp(\beta_2 \Delta z) - 1 = 0.11288$, expresado en porcentaje es 11.29% lo cual implica que para una unidad de incremento de esta variable se consigue una variación del 11.29% con respecto a la curva media del peso de la primera fase. En la Figura 30 se muestra la relación entre los residuos del primer modelado y la variable pH mínimo.

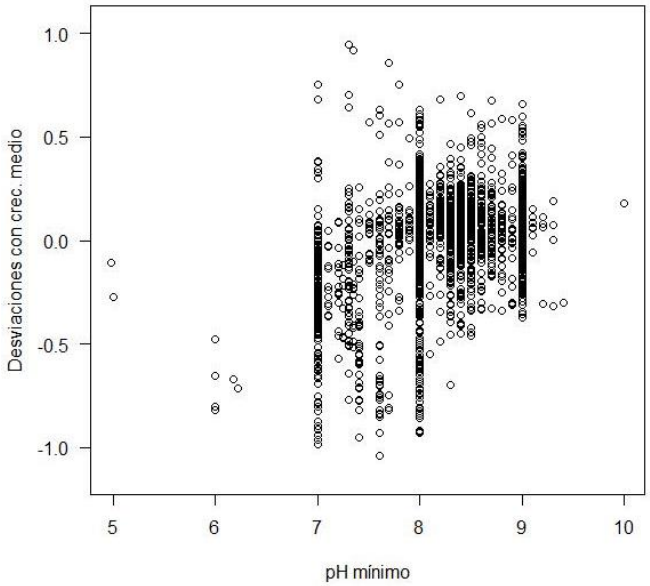


Figura 30. Gráfica de residuos vs pH mínimo.
Fuente: elaboración propia. Estimación en software R

Según la

Tabla 3, el coeficiente del oxígeno máximo semanal es -0.0172229. Tiene un efecto negativo con respecto al crecimiento esperado medio del langostino. Es decir, por cada unidad de incremento en el oxígeno máximo se espera una reducción de 0.0172229 en la curva media de la primera fase. Esto expresado en ratio del cambio de la variable $exp(\beta_3 \Delta z) - 1 = -0.01707543$, en términos de porcentaje es una reducción de la variación en 1.71% respecto de la curva media de la primera fase en la escala original del peso.

A pesar de que estos resultados parecen indicar que la relación entre oxígeno máximo y peso con respecto al crecimiento promedio es negativa, esta relación no es tan clara. Esto lo podemos observar en la Figura 31 se muestra la relación entre los residuos y la variable Oxígeno máximo.

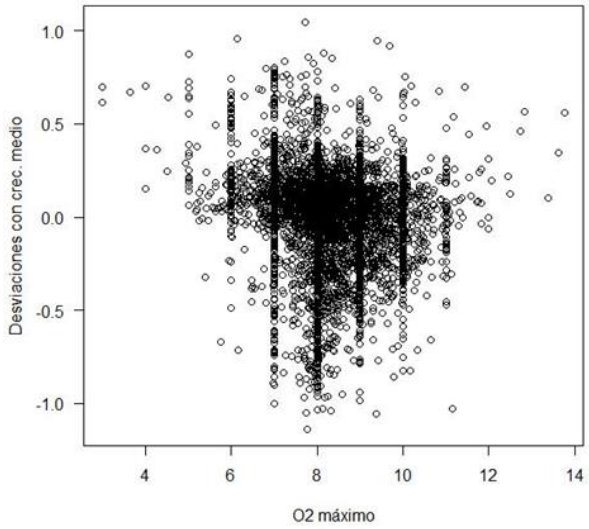


Figura 31. Gráfica de residuos vs O2 máximo.
Fuente: elaboración propia. Estimación en software R.

La temperatura máxima representa una contribución positiva al peso de los langostinos, esto es lógico debido a que los ambientes cálidos o temperaturas altas benefician al crecimiento y obtención de peso de los langostinos, de su coeficiente se puede inferir, que por cada unidad que se incrementa a la variable de temperatura se consigue un aumento medio de 0.0405461 en la curva del logaritmo del peso medio.

Este coeficiente representa un ratio de cambio de $exp(\beta_4 \Delta z) - 1 = 0.04137932$ lo cual en porcentaje es 4.14%, es decir, en un unidad de incremento en esta variable se obtiene una variación del 4.14% respecto a la curva media del peso. En la Figura 32 se muestra la relación entre los residuos y la variable Temperatura máxima.

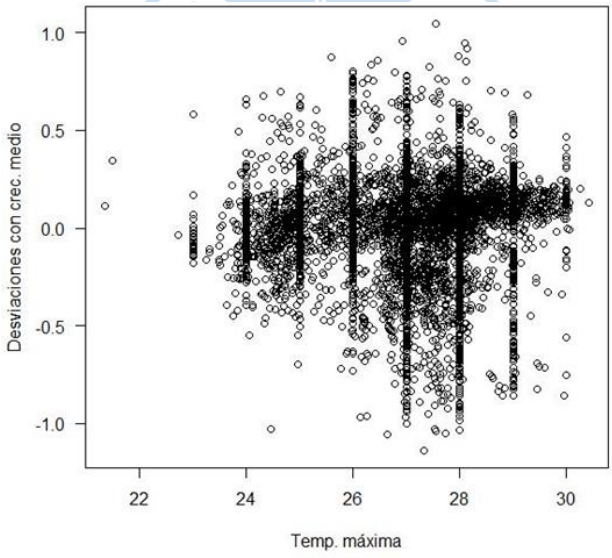


Figura 32. Gráfica de residuos vs Temp máxima.
Fuente: elaboración propia. Estimación en software R.

El coeficiente de la variable cianofitas (100K) es -0.0172375. Mantiene una relación negativa con el peso medio esperado del langostino, se puede decir que por cada unidad que se incremente la variable cianofitas (100K) se espera obtener una reducción de 0.0172375 en el logaritmo del peso del langostino.

Este resultado del coeficiente expresado como el ratio del cambio es $exp(\beta_5 \Delta z) - 1 = -0.01708978$ lo cual expresado en porcentaje es la reducción de la variación en 1.71% respecto del peso del langostino en la curva media de la primera fase en la escala de la variable peso. En la Figura 33 se muestra la relación entre los residuos y la variable cianofitas.

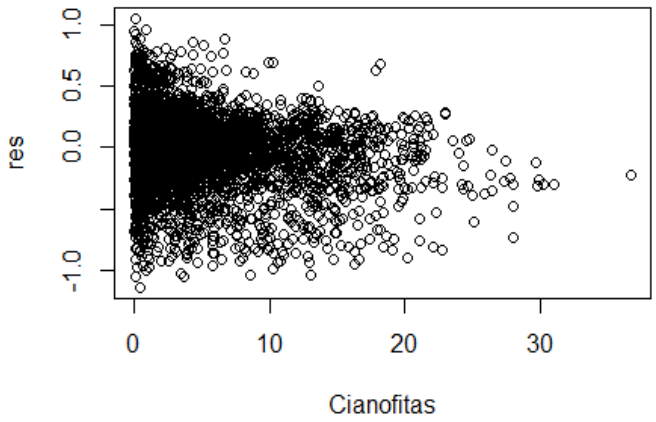


Figura 33. Gráfica de residuos vs cianofitas.
Fuente: elaboración propia. Estimación en software R.

El coeficiente de la variable alcalinidad es 0.0014434, esto sugiere que guarda una relación positiva con el peso medio esperado de los langostinos. Por cada unidad de incremento de la alcalinidad se espera que haya un aumento de 0.0014434 en el logaritmo del peso medio de la primera fase.

Este coeficiente equivale a un ratio de cambio de $exp(\beta_6 \Delta z) - 1 = 0.001444442$ por una unidad de variación de esta variable, que representa una variación del 0.14% respecto al peso medio del langostino en la curva de la primera fase en escala original del peso del langostino. Esta relación se puede explicar por la asociación entre la alcalinidad en las piscinas y el metabolismo de los langostinos. En la Figura 34 se muestra la relación entre los residuos y la variable alcalinidad.

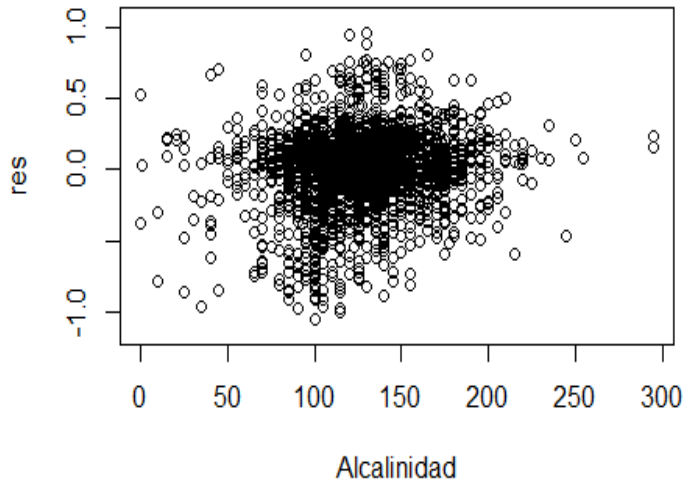


Figura 34. Gráfica de residuos vs alcalinidad.
Fuente: elaboración propia. Estimación en software R.

El factor de conversión acumulado tiene como coeficiente -0.1008640, es decir, esta variable tiene una relación negativa con el peso medio esperado de los langostinos. Se pude inferir que por cada unidad que se incremente a esta variable se espera una reducción de 0.1008640 al logaritmo del peso de los langostinos, lo que equivale a un ratio de cambio de $exp(\beta_7 \Delta z) - 1 = -0.09594402$ lo que en porcentaje es una reducción del 9.6%.

Según esto, para una unidad de variación de esta variable se obtiene una reducción en la variación respecto a la curva del peso medio de la primera fase en la escala original del peso del langostino. Esta inferencia puede sugerir la reducción del factor de conversión a cero para de esta manera evitar este efecto inverso, sin embargo, es necesario evaluar esto. En la Figura 35 se muestra la relación entre los residuos y la variable factor de conversión.

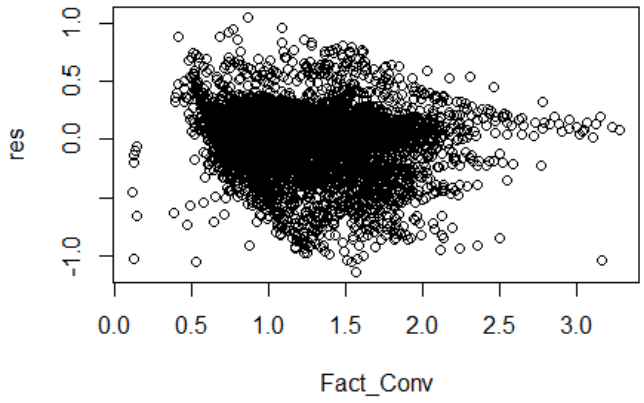


Figura 35. Gráfica de residuos vs Fac_conv.
Fuente: elaboración propia. Estimación en software R.

En cuanto a la variable categórica del tipo de piscina cuyo coeficiente es 0.1894017, lo que implica que existe un beneficio en la curva media del logaritmo del peso de los langostinos cuando las piscinas provienen de reservorios o son las piscinas P(Chi1) y P(Chi2); estas piscinas tienen algunas características o tratamiento diferenciador al resto.

Expresado este coeficiente como ratio de cambio es $exp(\beta_8 \Delta z) - 1 = 0.2085263$, es decir, representa un incremento en la variación del 20.85% respecto de la curva media del peso de los langostinos en la escala original.

El coeficiente de la variable categórica de mes de siembra es 0.1165539 para los meses que se encuentran dentro del grupo de mesC = "1" y 0.2192120 para los meses del grupo de mesC = "2". Según una interpretación aritmética, el primer coeficiente representa un aumento de 0.1165539 en relación a la curva media del logaritmo del peso del modelo no lineal.

El ratio de cambio es $exp(\beta_8 \Delta z) - 1 = 0.1236181$ lo cual representa en porcentaje de 12.36% de incremento de la variación respecto a la curva media del peso en la escala original. El coeficiente del segundo grupo presenta un incremento de 0.2192120 en la curva media del logaritmo del peso de la primera fase, en términos de ratio de cambio es $exp(\beta_8 \Delta z) - 1 = 0.2450952$ lo cual expresado en porcentaje representa un incremento de la variación del 24.51% respecto a la curva media del peso del langostino en la escala inicial.

Según las anteriores interpretaciones se podrían concluir, tomando como indicador el porcentaje de cambio, cuáles son las variables que generan un mayor impacto respecto del peso medio del langostino de entre estas variables. Sin embargo, se debe tener en cuenta que las variables difieren en la magnitud de sus dominios, por este motivo se necesita revisar estos porcentajes de cambio teniendo en cuenta esto.

En la Tabla 4 se muestran los porcentajes de cambio de cada una de las variables, como se muestra anteriormente, esta ratio se obtiene con la siguiente expresión $exp\ (\beta_n \Delta z) - 1$. A fin de visualizar el efecto del dominio en la proporción de cambio, se ha calculado las proporciones en relación a la variación máxima de su dominio, es decir, respecto a la variación del valor mínimo al valor máximo de cada una de las variables continuas.

Esa variación porcentual máxima se muestra en la última columna de la misma tabla. Por ejemplo, para el cálculo de la variación porcentual máxima de la variable predictora pH, dado que el rango de su dominio [4.98, 10], el máximo aumento de dicha variable es de 5.02 (si es inicialmente el valor de dicha variable es de 4.98, asumiendo que las demás variables permanecen constantes). Para dicha situación, la variación máxima porcentual en ϵ_{ij} (variación del peso con respecto al peso medio) sería, según esta expresión $exp(0.1069514 * (5.02)) - 1$, 71.07%.

Tras realizar las anteriores precisiones se puede reconocer el máximo impacto de las variables teniendo en cuenta su rango, y qué tanto impacto pueda tener los cambios de estas variables en el crecimiento medio del langostino.

Tabla 4. Variación porcentual en ϵ_{ij} (variación peso con respecto al peso medio) por una unidad de cambio en las variables predictoras y variación máxima.

	Dominio	Δ% en ϵ_{ij}	Δ% máxima en ϵ_{ij}
Nicovita	[0, 1]	4.21	4.213126
pH_min	[4.98, 10]	11.29	71.068868
O2_max	[3, 13.77]	-1.71	-16.932341
T_max	[21.4, 30.4]	4.14	43.497606
Cianofitas	[0, 36.725]	-1.71	-46.902864
Alcalinidad	[0.06, 295]	0.14	52.993560
Fact_Conv	[0.12, 3.28]	-9.6	-25.318647

Fuente: elaboración propia. Estimación en software R.

Capítulo 4

Análisis de sensibilidad y bondad de ajuste del modelo

En este capítulo se analiza y evalúa la sensibilidad de los coeficientes y la bondad de ajuste del modelo que resultó de la etapa de modelamiento. Para realizar este análisis se hará uso de una validación cruzada con ayuda de un algoritmo que permita agrupar la data aleatoriamente y realizar una serie de operaciones repetitivas.

4.1. Validación cruzada "*Hold out*"

Es un método de remuestreo de datos que permite cuantificar la incertidumbre asociada con algún estimador estadístico como el error del modelo y los parámetros de ajuste del modelo. Existen diferentes técnicas para realizar esta validación, pero para este estudio nos centraremos en el método "*Hold out*" el cual está compuesto por 3 etapas. La primera es el entrenamiento de la data, la segunda es la validación y la tercera es la prueba. Para el caso particular de este estudio, la etapa de validación se ha hecho previamente en el capítulo de modelamiento donde se determinó el modelo no lineal de la primera fase y las variables significativas del modelo lineal de la segunda fase.

Esta técnica denomina "*data learning*" a toda la data que se usó para la construcción del modelo, y plantea la creación de dos subconjuntos a partir de esta data. El primer subconjunto denominado "*data train*" que hace referencia a la data de entrenamiento y el segundo subconjunto "*data test*" a la data de prueba, datos que no son utilizados en el entrenamiento. (Berrar , 2018)

Estos subconjuntos se crean de manera que aleatoriamente cierto porcentaje de la *data learning* se destina a cada uno de ellos. La *data train* se utiliza para la creación de un modelo que posteriormente se evalúa con ayuda de la *data test* haciendo la estimación por medio de este modelo y contrastándolo con los valores reales de este último subconjunto. De esta manera se puede observar el grado de variabilidad del modelo con respecto a la variación de la data usada para su construcción y la bondad del ajuste frente a data distinta.

Esta agrupación en subconjuntos se realiza "n" veces, es decir, que se tendrán "n" subconjuntos de *data train* con los que se construye un modelo diferente para cada uno. Los valores de los coeficientes estimados, generados para cada uno de estos modelos construidos, se utilizan para evaluar con ayuda de los valores de la *data test*, subconjunto complementario correspondiente, los valores estimados versus los valores observados. Al ser una tarea repetitiva es necesario la programación de un algoritmo para que realice las operaciones tantas veces como se determine.

4.2. Procedimiento

Dado que se ha usado la data completa para la estimación del modelo de dos fases, se ha querido evaluar qué tan sensible son los valores estimados de los coeficientes de dichas variables usando solo parte de la data para la estimación. Si este proceso se repite un número de veces, siempre seleccionando aleatoriamente parte de la data, podríamos obtener una idea de qué tan sensible son esos estimados con respecto a la data utilizada de dicha campaña.

El algoritmo utilizado se puede describir brevemente de la siguiente manera:

- Paso 1. Aleatoriamente se escogen el 70% de las piscinas para la estimación del modelo de dos fases mostrado en el capítulo 3.
- Paso 2: Se guarda el coeficiente estimado de las variables explicativas del modelo de regresión lineal múltiple.
- Paso 3. Repetir el proceso K veces. En este estudio K=10000.

Con la *data train*, el primer 70% de la *data learning*, se construye el modelo de 2 fases. Primero, el modelo no lineal según el modelo *Von Bertalanffy*, se almacenan los coeficientes determinados y los residuos. Segundo, a partir de los residuos almacenados se construye el modelo lineal multivariable, incluyendo solo las variables definidas en el capítulo anterior.

A partir de estos modelos construidos, realizamos la prueba con la *data test*, el 30% de la *data learning* restante. Esto consiste en utilizar los coeficientes de los modelos, construidos anteriormente, para calcular la estimación mediante el ingreso de los datos de la *data test*. Es decir, se realizan las mismas operaciones de estimación correspondientes para ambas fases que se usaron en el modelamiento del capítulo anterior, pero ya no se calculan los modelos, sino se ingresan las variables de entrada correspondientes al modelo y se estima el resultado.

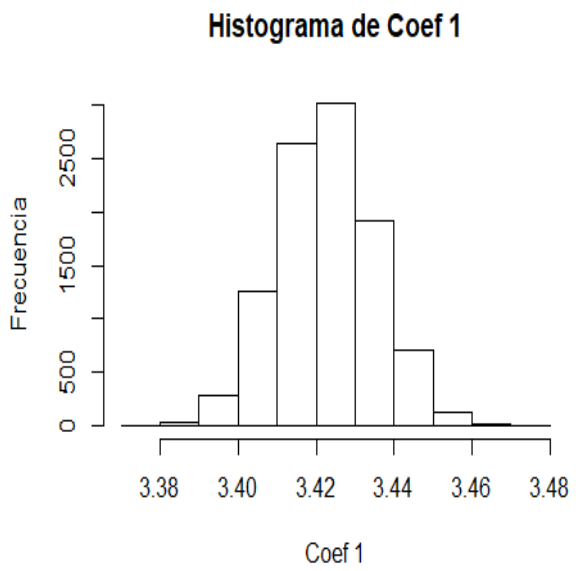
Estas estimaciones se comparan con las observaciones reales para obtener los indicadores correspondientes, de manera que podamos analizar la sensibilidad de los modelos, frente a data diferente de la usada para su construcción, y la variabilidad de sus coeficientes.

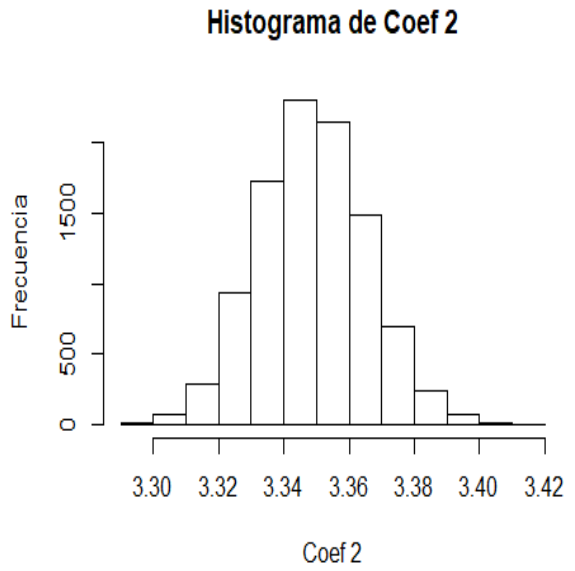
4.3. Análisis de sensibilidad del modelo no lineal

Para el modelo no lineal se calculan los coeficientes del modelo a partir de la *data train*, luego ingresando la *data test* se estiman los valores de la variable respuesta y al compararse con las observaciones reales se obtiene la bondad de ajuste del modelo y el error cuadrático medio.

En base a los coeficiente calculados durante las 10000 réplicas se construyen los histogramas que se muestran en la Figura 36, donde se observa que la distribución de los valores de los coeficiente siguen una distribución normal, tal como lo indica la teoría bajo la hipótesis de normalidad que los estimados de coeficientes de modelos no lineales deben tener.

Se observa su unimodalidad, baja dispersión y aparente simetría con respecto a los valores medios; estos valores se encuentran muy cercanos a los valores obtenidos en el modelamiento del capítulo anterior. Esta baja variabilidad entre los coeficientes estimados se evidencia en los bajos valores de varianza mostrados en la Tabla 5, que se presenta el resumen estadístico de estos. Es decir, existe una estabilidad de los coeficientes frente a la variación de los datos utilizados para la construcción de los modelos.





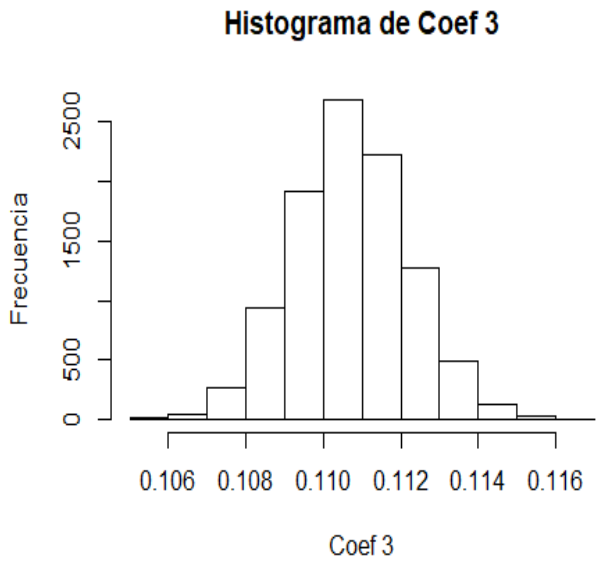


Figura 36. Histogramas de los coeficientes estimados de las variables del modelo no lineal.
Fuente: elaboración propia.

Tabla 5. Resumen estadístico de los valores estimados de los coeficientes del modelo no lineal.

	Media	Mediana	Desv. Estandar	Varianza
Coef 1	3.4226023	3.4224269	0.0124838	0.0001558
Coef 2	3.3487452	3.3486808	0.0166216	0.0002763
Coef 3	0.1107051	0.1106954	0.0014806	0.0000022

Fuente: elaboración propia.

Se analiza, como se mencionó anteriormente, la variación de la bondad de ajuste y el error cuadrático medio al variar el grupo de datos para la construcción del modelo no lineal. La evaluación de estos 10000 modelos construidos se realiza al contrastar la estimación del resultado, a partir de los valores de la *data test,* con lo valores observados de la misma data.

Los histogramas de los resultados obtenidos se muestran en la Figura 37 donde se observa un claro comportamiento unimodal, baja dispersión y simietría lo cual supone una normalidad de estos. La mayoría de los modelos construidos han tenido una bondad de ajuste que varía entre el 0.8 y 1, y con unos valores de MSE conprendidos, en su mayoría, entre 0.006 y 0.011.

Según la Tabla 6, el valor promedio aproximado del R cuadrado ajustado es 0.9031933, lo cual supone una muy buena representatividad del modelo no lineal, mientras que el bajo valor medio del MSE refuerza la idea que el modelo estimada bien el crecimiento medio. Estos indicadores que se muestran en la Tabla 6 se encuentran asociados a valores bajos de varianza.

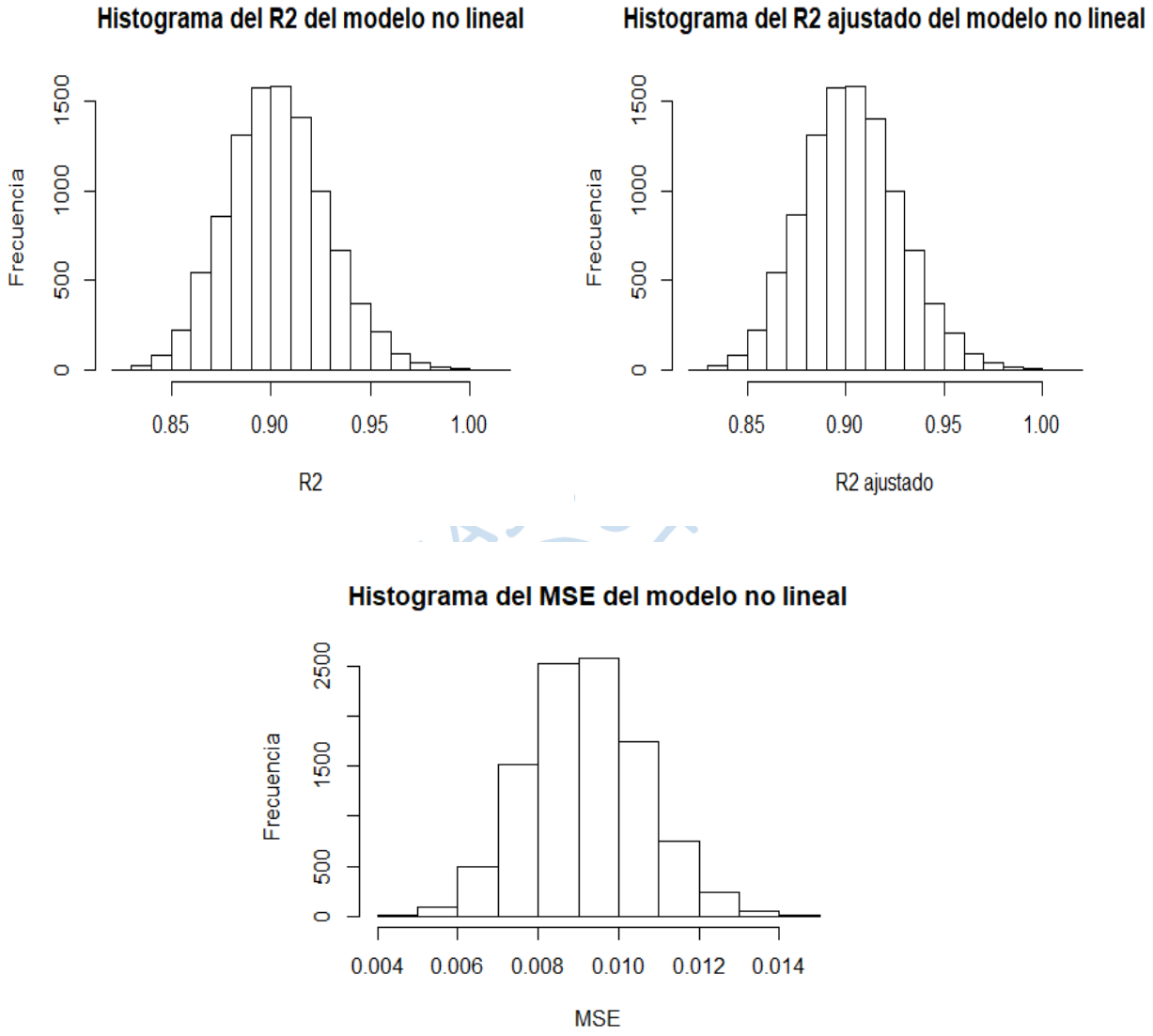


Figura 37. Histogramas de los indicadores estadísticos estimados del modelo no lineal.
Fuente: elaboración propia.

Tabla 6. Resumen estadístico de los valores estimados de los indicadores estadísticos del modelo no lineal.

	Media	Mediana	Desv. Estándar	Varianza
R2	0.9032487	0.9023917	0.0246777	0.0006090
R2a	0.9031933	0.9023365	0.0246916	0.0006097
MSE	0.0091889	0.0091511	0.0014446	0.0000021

Fuente: elaboración propia.

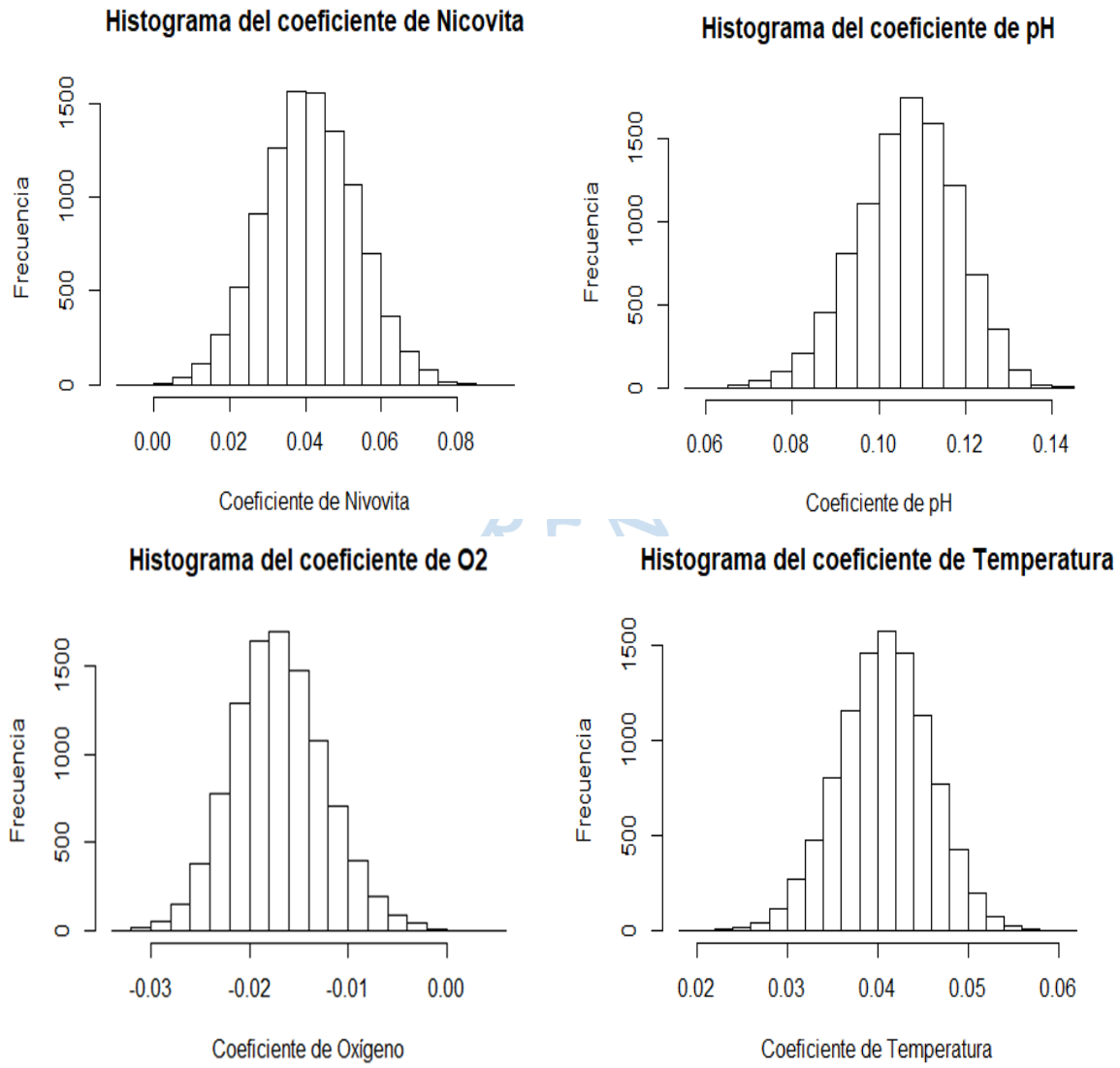
A partir de los anteriores resultados se puede suponer que los valores, de la primera fase del modelo, no varían dramáticamente con la variación de los datos usados para la construcción de este, es decir, los coeficientes del modelo no lineal, así como, su R cuadrado, R cuadrado

ajustado y MSE mantienen de cierta manera una estabilidad que asegura la representatividad del crecimiento medio del langostino expresado como el logaritmo del peso en el tiempo.

4.4. Análisis de sensibilidad del modelo lineal

Para el análisis de sensibilidad de la segunda fase se utilizan los residuos de la anterior fase y se construye el modelo lineal múltiple. La Figura 38 muestra los histogramas de los valores de los coeficientes estimados de cada una de las variables del modelo lineal múltiple. Se puede observar que estos valores, tal como la teoría lo indica para los estimadores de los coeficientes de modelos lineales, siguen una distribución normal debido a su forma unimodal centrada y con una baja dispersión.

En la Tabla 7 se muestra la media, mediana y desviación estándar de los 10000 valores de cada uno de los coeficientes estimados. Este resumen estadístico nos muestra bajos valores de varianza respecto al valor medio lo cual supone una baja variabilidad de los coeficientes.



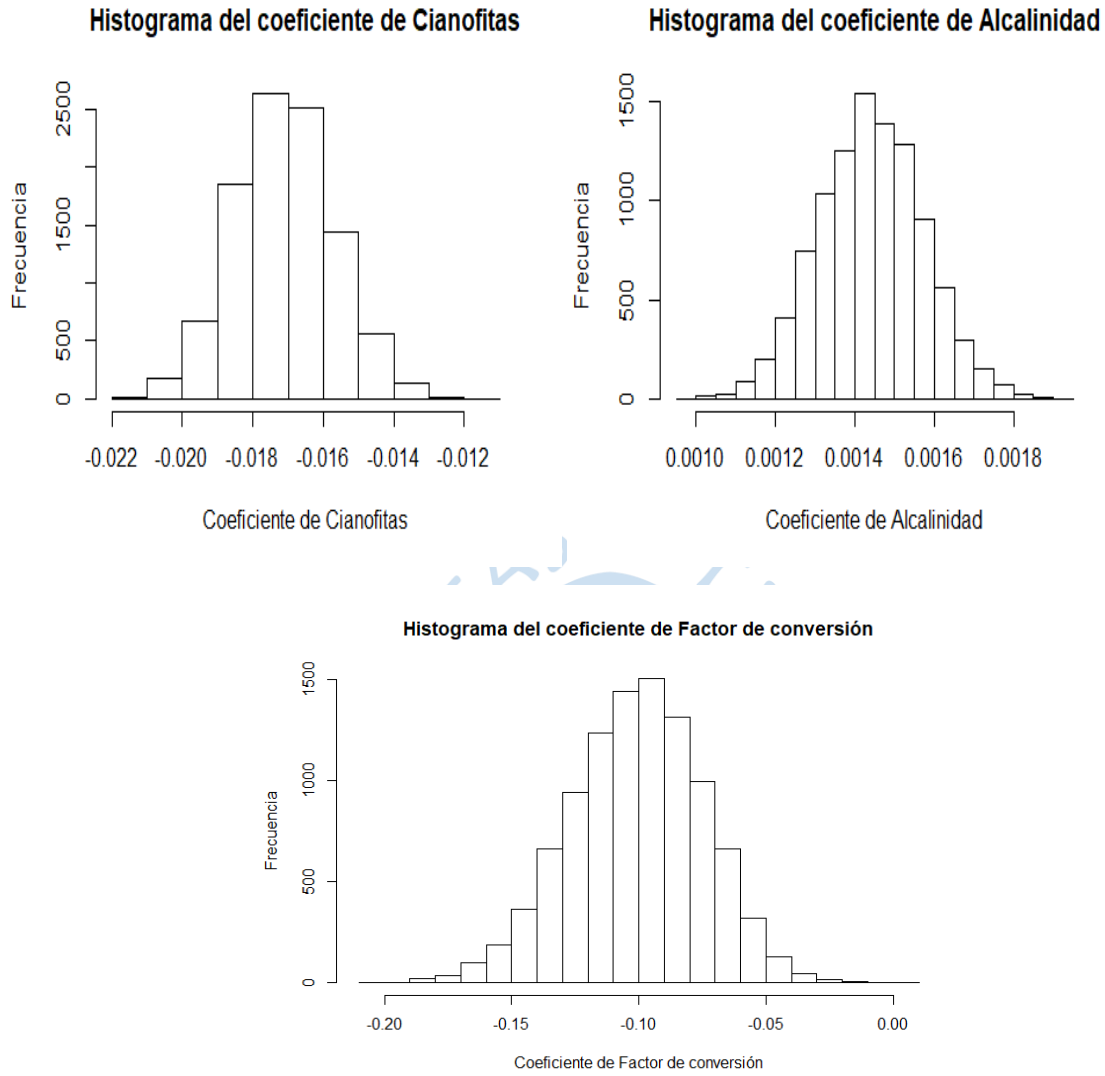


Figura 38 . Histogramas de los valores estimados de las variables del modelo lineal.
Fuente: elaboración propia.

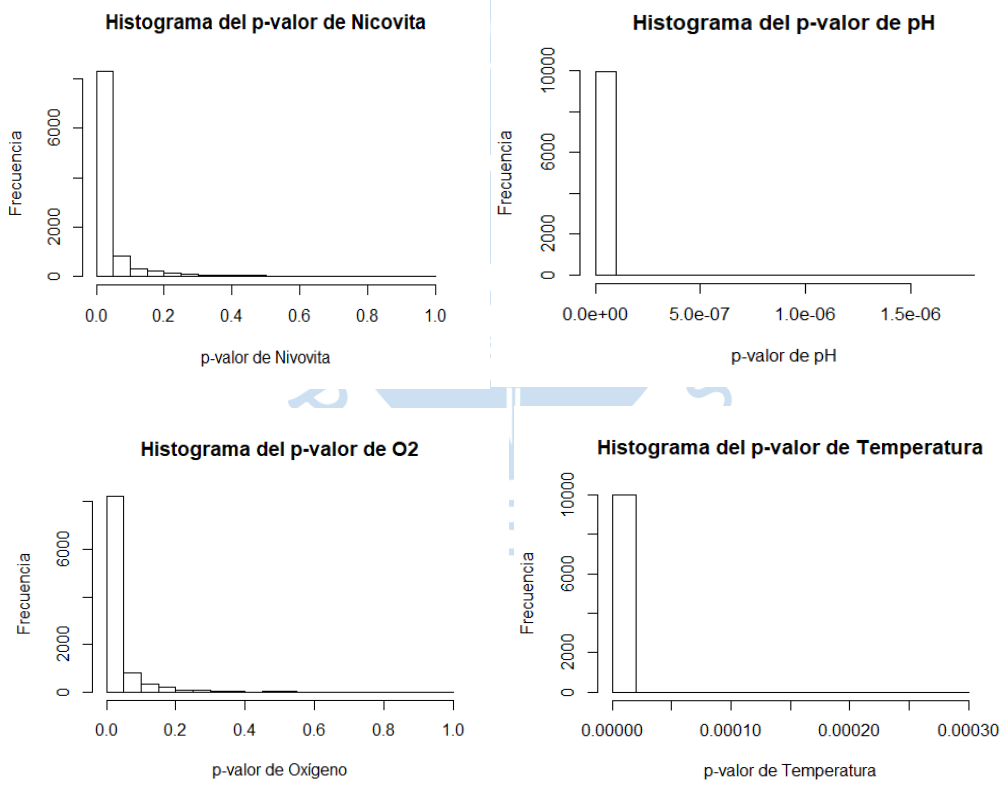
Tabla 7. Resumen estadístico de los valores estimados de los coeficientes de las variables.

	Media	Mediana	Desv. Estándar	Varianza
Nicovita	0.0411491	0.0409407	0.0124096	0.0001540
pH	0.1065191	0.1071185	0.0115801	0.0001341
O2	-0.0169783	-0.0171815	0.0047113	0.0000222
Temperatura	0.0407422	0.0408255	0.0050458	0.0000255
Cianofitas	-0.0171077	-0.0171298	0.0014129	0.0000020
Alcalinidad	0.0014407	0.0014405	0.0001315	0.0000000
Factor Conversión	-0.1009004	-0.0999959	0.0263154	0.0006925

Fuente: elaboración propia.

La construcción de diferentes modelos permite evaluar la variabilidad de los coeficientes de las cada una de las variables, pero también permite analizar la sensibilidad de la relevancia de estas variables en el modelo. Para realizar esta evaluación se almacenaron también los p-valores de las variables en cada uno de los modelos creados. Los resultados obtenidos se presentan en los histogramas de la Figura 39 donde se puede observar que estos se encuentran concentrados en valores muy cercanos a cero.

De entre todas las variables, nicovita y oxígeno tienen una variabilidad más notable en la relevancia para el modelo lineal es por esta razón que presenta una ligera asimetría positiva. Sin embargo, como se muestra en la Tabla 8, todos los p-valores medios de las variables se encuentran por debajo del 0.05, valor de significancia usado. Además, de estar acompañados de una varianza relativamente baja. Es decir, a pesar de la ligera sensibilidad de las dos variables mencionadas anteriormente, estas se han encontrado como significativas en la gran mayoría de los modelos construidos junto con las demás variables.



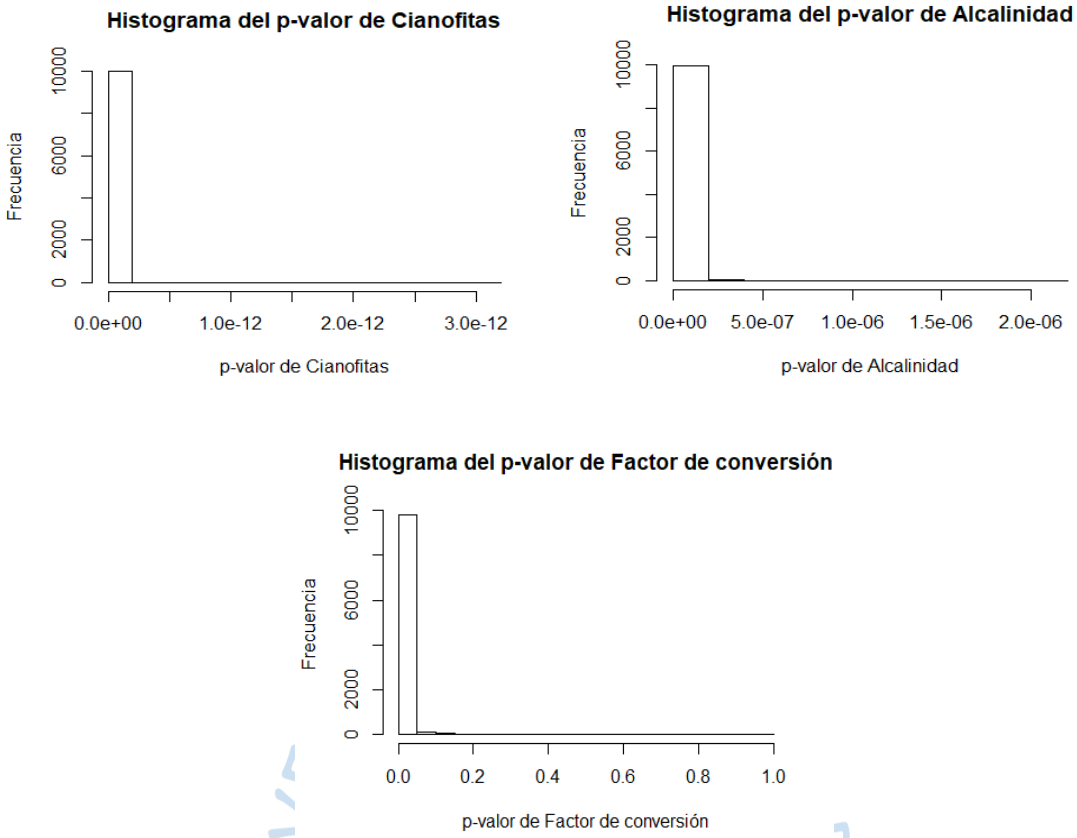


Figura 39. Histogramas de los p-valores estimados de las variables del modelo lineal.
Fuente: elaboración propia.

Tabla 8. Resumen estadístico de los p-valores estimados para una de las variables del modelo lineal.

	Media	Mediana	Desv. Estándar	Varianza
Nicovita	0.0331083	0.0061408	0.0768190	0.0059012
pH	0.0000000	0.0000000	0.0000000	0.0000000
O2	0.0360828	0.0078899	0.0817138	0.0066772
Temperatura	0.0000001	0.0000000	0.0000035	0.0000000
Cianofitas	0.0000000	0.0000000	0.0000000	0.0000000
Alcalinidad	0.0000000	0.0000000	0.0000001	0.0000000
Factor Conversión	0.0044638	0.0000761	0.0248695	0.0006185

Fuente: elaboración propia.

Para evaluar la sensibilidad del modelo lineal múltiple también es necesario analizar la variabilidad de su bondad de ajuste y error cuadrático medio al igual como se hizo con la primera fase del modelo. En la Figura 40 se muestran los histogramas de dichos indicadores, las gráficas se muestran bastante concentradas a sus valores medios y con una baja amplitud, es decir, presentan un comportamiento normal debido a su unimodalidad, simetría y relativa poca

dispersión. En la Tabla 9 se muestra el resumen estadístico donde se observan bajos valores de varianza para el Rcuadrado, Rcuadrado ajustado y el MSE esto hace suponer que la variabilidad causada por la data que se utilice para la construcción del modelo es relativamente baja. Además de obtenerse un R cuadrado ajustado de 0.51 lo cual a simple vista parece muy bajo, sin embargo, es importante para fines de inferencia.

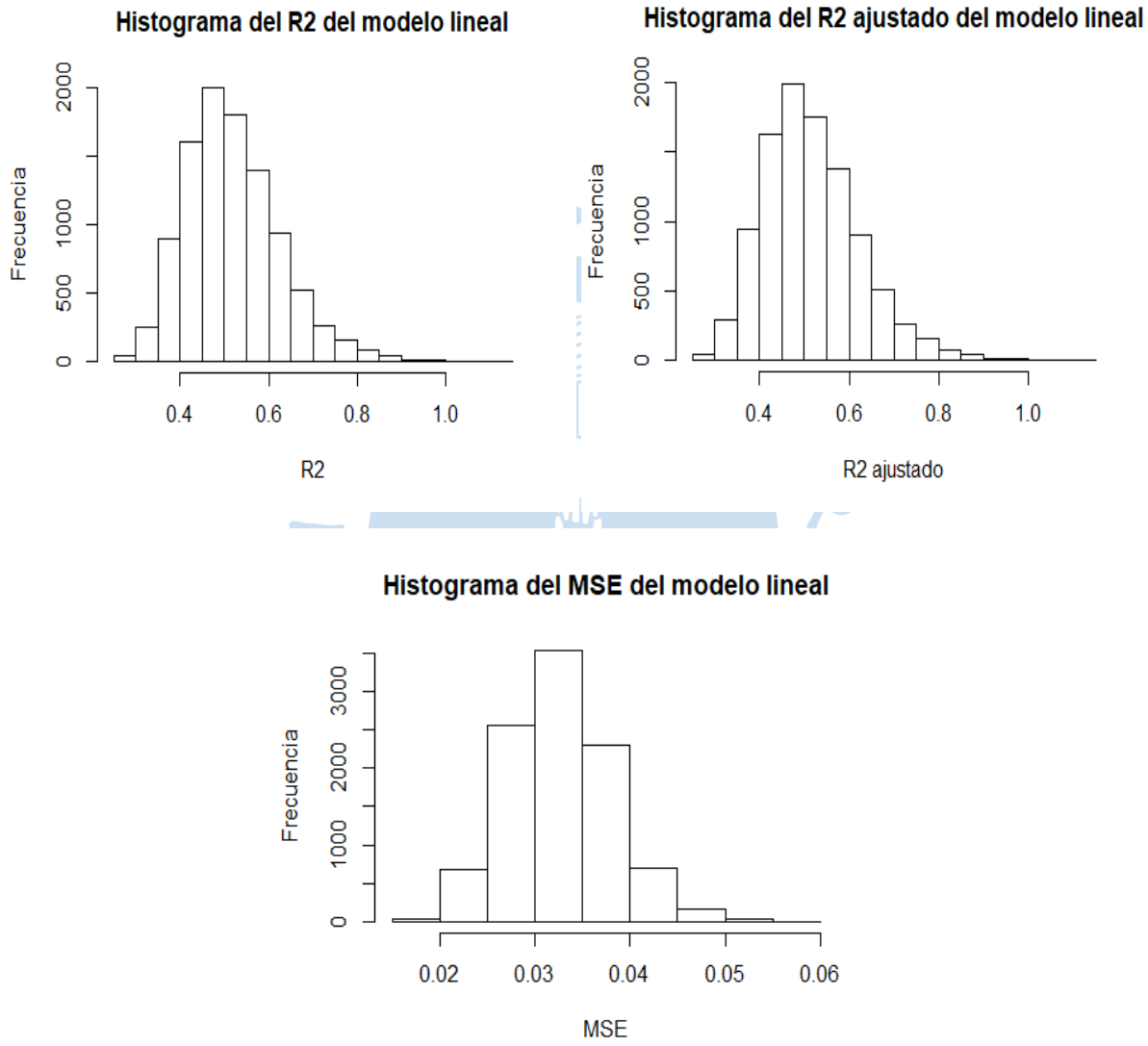


Figura 40. Histogramas de los valores estimados R cuadrado, R cuadrado ajustado y el MSE.
Fuente: elaboración propia.

Tabla 9. Resumen estadístico de los valores estimados de los indicadores estadísticos del modelo lineal.

	Media	Mediana	Desv. Estándar	Varianza
R2	0.5178982	0.5052060	0.1066270	0.0113693
R2a	0.5154011	0.5026786	0.1071810	0.0114878
MSE	0.0326338	0.0323880	0.0054353	0.0000295

Fuente: elaboración propia.

Este análisis permite reconocer la variabilidad del nivel de representatividad del modelo no lineal según la data con la que se trabaje. El modelo presenta un valor muy bueno de bondad de ajuste el cual presenta un comportamiento bastante estable lo que significa que el modelo se ajusta muy bien a datos diferentes, dentro de la campaña usada. Esta baja variabilidad asegura el aislamiento del crecimiento medio esperado del langostino de los diferentes factores que afectan al crecimiento de este, y también, que los resultados obtenidos por este procedimiento son aplicables a otra data similar.

A partir de la validación anterior se analiza la sensibilidad del modelo lineal múltiple, siendo esta la etapa más importante de este estudio porque es aquí donde se identifican las variables significativas. El análisis de sensibilidad nos muestra que los coeficientes mantienen sus valores de manera estable a pesar del uso distinto de la data, dentro de una misma campaña. Cabe resaltar que los coeficientes que se determinaron en la etapa de modelamiento se encuentran contenidos dentro de los valores estimados para los modelos construidos. Esta estabilidad de los coeficientes es acompañada por la rigidez de la bondad de ajuste y su MSE.

Además, el análisis de sensibilidad también permite reconocer la variabilidad de la significatividad de las variables que se determinaron previamente. El resultado es que todas las variables incluidas en el modelo, en media, son significativas aun cuando la data es diferente. Este resultado respalda que las variables determinadas sí tienen un impacto significativo en la curva de crecimiento del langostino.

Según lo mencionado anteriormente la sensibilidad de ambas fases del modelo tiene una variabilidad esperada, es decir, cuenta con una baja variabilidad generada por la evaluación del modelo en data diferente a la de su construcción. Sin embargo, dicha baja variabilidad respalda que el modelamiento por dos fases es lo suficientemente representativo y estable como para utilizarse para determinar las variables de mayor influencia en el crecimiento de los langostinos, para esta campaña de cultivo en la empresa.

UNIVERSITAS STUDIORUM PIURENSIS

Conclusiones y recomendaciones

La limpieza de los datos es una actividad crítica para el análisis, a causa del origen de la fuente de los datos y su gestión, es necesario no solo detectar los errores, sino clasificarlos y tomar las correspondientes medidas correctivas, para poder realizar un correcto análisis estadístico y por ende obtener conclusiones razonables y que ayuden a mejorar el proceso.

Los errores también son muy importantes porque pueden entregar información útil sobre algún tratamiento anormal, evento atípico o alguna mala práctica con el tipeo y almacenamiento de los datos que cuya detección puede ayudar a entender mejor el proceso o los cuidados necesarios que se le debe dar a la información que se levante.

Es muy importante seguir un proceso ordenado de identificación y corrección de los errores en la data, para asegurar que se realice un correcta y ordenada limpieza. Adecuar la estructura de los datos para que puedan ser procesados por los softwares, definir las herramientas estadísticas que permitirán identificar a los datos atípicos, establecer una clasificación para relacionarlos con las fuentes que los podrían generar y determinar las medidas correctivas más convenientes.

La importancia del modelamiento radica no solo en los resultados de predicción, sino también, en la interpretabilidad de las relaciones de las variables. En este último punto es en el que se resalta la importancia de este estudio debido a que contribuye a la identificación de aquellas variables críticas para el crecimiento de los langostinos, tras un análisis lógico que lo sustente.

El modelamiento por dos fases permite separar el comportamiento tendencial y esperado medio del crecimiento respecto del tiempo, y analizar de manera “aislada” los ruidos, es decir, aquellas variables que podrían tener un efecto significativo en el desarrollo de los langostinos, cuya identificación sería mucho más complicada sino se hiciera este trabajo previo.

Las variables que resultan ser más importantes son: la temperatura, el oxígeno, pH, alcalinidad, nicovita, cianofitas, factor de conversión acumulado y las variables categóricas del tipo de piscina, del mes de siembra. Estas son las variables identificadas como independiente (sin

colinealidad), que explican de forma significativa la diferencia entre el peso del langostino y el crecimiento medio de los langostinos. Por lo tanto, las variables que describen las condiciones de la piscina de siembra, alimentación, cantidad de alimento acumulado por animal, mes de siembra, y el tipo de piscinas afectan en cierta medida al crecimiento de los langostinos, y son en ellas en las que se deben enfocar la atención para monitorear y posiblemente mejorar el proceso.

El uso de las transformaciones en las variables es muy utilizado en los modelamientos. En el caso de la transformación logarítmica que se aplicó a la variable respuesta con la finalidad de tener un comportamiento progresivo más uniforme y que evidencie el comportamiento tendencial del peso en el tiempo, también facilitó la interpretación de los coeficientes del segundo modelo lo cual es uno de los objetivos del presente trabajo luego de la identificación de las variables críticas.

Los coeficientes de las variables independientes, identificadas como importantes en este proceso, permiten la interpretación de estas relaciones. Por ejemplo, la temperatura según su coeficiente representa una contribución positiva a la curva media del logaritmo al peso del modelo de la primera fase, sin embargo gracias a la interpretación geométrica que permite la transformación inicial de la variable respuesta, se puede identificar cuál es el impacto en la variación del peso medio en la escala sin la transformación, de esta manera se puede reconocer la importancia de cada una de las variables en el proceso de crecimiento de los langostino.

El análisis de sensibilidad por validación cruzada permitió reconocer qué tan estable son los resultados obtenidos por el modelamiento de dos fases, según la data con la que se construyan los modelos frente a los datos con los que se evalúe. Se concluye, entonces, que el modelamiento por dos fases es bastante bueno para fines interpretativos, como lo ha sido para la identificación de las variables más importantes en crecimiento de los langostinos, del presente trabajo.

Los resultados obtenidos, por el modelamiento por dos fases, están condicionados por la data utilizada y muchas otras variables que no se han identificado y medido, por ello no hay datos sobre estas, si bien porque se desconocen o son muy difíciles de medir. Además, la estadística es una herramienta que nos permite guiar acciones y decisiones, sin embargo, debe ir acompañada de la opinión de expertos. En el caso de los resultados de este estudio han sido validados por el personal relacionado al proceso, quienes consideran a la temperatura, oxígeno, pH y alcalinidad como variables críticas para el crecimiento de los langostinos. En los resultados este trabajo identifica a todas estas variables junto con otras como significativas para el crecimiento del langostino, alineándose a la opinión de los expertos.

Referencias bibliográficas

Betancourt, G. (2005). *Scientia et Technica Año XI.* Risaralda: Universidad Tecnológica de Pereira.

Aggarwal, C. (2013). *Outliers Analysis.* New York: IBM Thomas J. Watson Research Center.

Amón Uribe, I. (2010). *Guía de metodología para la selección de técincas de depuración de datos.* Medellín: Universidad Nacional de Colombia.

Araneda, M., Hernández, J., Gasca-Leyva, E., & Vela, M. (2013). Growth modelling including size hetrogeneity: Application to the intensitive culture of white shrimp (P. vannamei) in freshwater. *Elsevier*, 1-12.

Ben-Gal, I. (2005). *eng.tau.ac.il.* Fonte: Department of Industrial Engineering Tel-Aviv University: http://www.eng.tau.ac.il/~bengal/outlier.pdf

Berrar , D. (2018). *researchgate.net.* Fonte: https://www.researchgate.net/publication/324701535_Cross-Validation

Caballero Solano, S. M. (2015). *Análisis de costos de cultivo de larvas de camarón y su influencia en la rentabilidad de la empresa Maramar S.A.* Guayaquil: Universidad Politécnica Salesiana. Fonte: https://dspace.ups.edu.ec/bitstream/123456789/9994/1/UPS-GT001132.pdf

Carbajal Hernández, J., Sánchez-Fernández, L., & Villa Vargas, L. (Diciembre de 2012). *researchgate.net.* Fonte: https://www.researchgate.net/publication/254258523_Water_quality_assessment_in_shrimp_culture_using_an_analytical_hierarchical_process

Casas, G., Rodríguez, D., & Afanador Téllez, G. (2010). Propiedades matemáticas del modelo Gompertz y su aplicación del crecimiento de los cerdos. *Revista Colombiana de Ciencias Pecuarias*, 349-358. Fonte: https://www.redalyc.org/pdf/2950/295023477010.pdf

Chandola, V., Banerjee, A., & Kumar, V. (2009). *Anomaly Detection: A Survey.* Minneapolis: University of Minnesota.

Científica, F. d. (8 de Diciembre de 2011). *estadistica-dma.ulpgc.es.* Fonte: El modelo de crimiento de Von Bertalanffy: http://www.estadistica-dma.ulpgc.es/FCC/pdf/Bertalanffy.pdf

ComexPerú. (28 de setiembre de 2018). *comexperu.org.pe*. Fonte: https://www.comexperu.org.pe/articulo/impulso-para-la-acuicultura-aunque-pudo-ser-mejor

D. Angrist, J., & Pischke , J.-S. (2015). *Mastering Metrics.* Nueva Jersey: Princeton University Press.

De Armas, A. A. (2015). *Detección de Outliers en grandes bases de datos.* Buenos Aires: Universidad Argentina de la Empresa.

Departamento de inteligencia de Mercados. (2017). *PromPerú.* Fonte: http://www.siicex.gob.pe/siicex/documentosportal/773608272radC6745.pdf

Díaz Herrera, F., Juárez Castro, G., Pérez Cruz, E., & Bückle Ramírez, F. (1991). *Balance energético de postlarvas y juveniles del langostino Malayo macrobrachium rosenbergü de man (crustacea:palaemonidae).* Ciudad de México: Universidad Autónoma de México.

Echenique, R., & González, D. (1998). *http://sedici.unlp.edu.ar/.* Fonte: http://sedici.unlp.edu.ar/bitstream/handle/10915/49319/Documento_completo__.pdf?sequence=1&isAllowed=y

ECOSAC. (2015). *academia.edu/.* Fonte: https://www.academia.edu/37144373/ECOSAC_TRABAJO_FINAL?email_work_card=view-paper

Equilibrium. (2019). *bvl.com.pe.* Fonte: https://www.bvl.com.pe/hhii/L00251/20190618165701/INFORME322.PDF

Espino Timón, C. (2017). *Análisis predictivo: técnicas y modelos utilizados y aplicaciones del mismo - herramientas Open source que permiten su uso.* Catalunya : Universitat Oberta de Catalunya.

FAO. (2009). *fao.org.* Fonte: http://www.fao.org/tempref/FI/DOCUMENT/aquaculture/CulturedSpecies/file/es/es_whitelegshrimp.htm

FAO. (2018). *El estado mundial de la pesca y la acuicultura.* FAO publications. Fonte: fao.org: http://www.fao.org/3/I9540ES/i9540es.pdf#page=19

FAO. (Noviembre de 2019). *Organización de las Naciones Unidas para la Alimentación y la Agricultura*. Fonte: fao.org: http://www.fao.org/aquaculture/es/

Gareth, J., Witten, D., Hastie, T., & Tibshirani, R. (2017). *An Introduction to Statistical Learning*. New York: Springer.

Goodall, E., & Sprevak, D. (1984). *A note on a Stochastic model to describe the milk yield of a dairy cow*. Anim Prod.

Guevara, A., & L. C. (2015). *Dinámica de los grupos de fitoplancton Clorófita y Cianofita, y su relación con los parámetros físico-químicos en las pilas de estabilización San Isidro.* Managua: Universidad Nacional Autónoma de Nicaragua. Fonte: file:///C:/Users/To%C3%B1o/Downloads/guevara,%20Calix.pdf

Guzman Valqui, B. K., & Leiva Tafur, D. (2015). *Uso de diatomeas (Bacillariophyceae) como bioindicadores para la evaluación de la calidad del agua en la cuenca del río Utcubamba.* Amazonas: Universidad Nacional "Toribio Rodríguez de Mendoza de Amazona".

Hassan Darmani kuhi, S. L. (Junio de 2010). *A review of mathematical functions for the analysis of growth in poultry.* Fonte: pdfs.semanticscholar.org: https://pdfs.semanticscholar.org/369d/85d85c8ed39334f8e5430d71898df45660dd.pdf

Hawkins, D. (1980). *Identification of Outliers (Monographs on Statistics and Applied Probability) vol 3.* Londres: Chapman and Hall.

INEI. (Febrero de 2019). *Inei.gob.pe.* Acesso em 18 de Octubre de 2019, disponível em https://www.inei.gob.pe/media/principales_indicadores/informe-tecnico-de-produccion-nacional-febrero2019.PDF

Knorr, E. M. (2002). *Outliers and Data Mining: Finding Exceptions in Data.* Vancouver: The University of British Columbia.

Larrañaga, P., Inza, I., & Moujahid, A. (2000). *sc.ehu.es.* Fonte: http://www.sc.ehu.es/ccwbayes/docencia/mmcc/docs/t6bayesianos.pdf

León Caminati, A. R. (2017). *Sistema de monitoreo de variables críticas en el proceso productivo de cultivo de langostino en agua dulce.* Piura: Universidad de Piura.

López Hernández, J. A., & Valles Gándara, A. G. (Junio de 2009). *Modelos para la estimación del índice de sitio para Pinus durangesis Martínez en San Dimas, Durango.* Fonte: http://www.scielo.org.mx/pdf/cfm/v34n105/v34n105a10.pdf

Medina, F., & Galván, M. (2007). *Imputación de datos: teoría y práctica.* Santiago de Chile: Naciones Unidas (CEPAL).

Ministerio de la producción. (2017). *Produce.gob.pe.* Fonte: http://www2.produce.gob.pe/RepositorioAPS/3/jer/ACUISUBMENU4/boletines/DESCRIPCI%C3%93N%20DE%20LA%20ACTIVIDAD%20ACUICOLA.pdf

Mishra, A., Verdegem, M., & Van Dam, A. (Diciembre de 2001). *researchgate.net.* Fonte: https://www.researchgate.net/publication/40141237_A_Dynamic_Simulation_Model_for_Growth_of_Penaeid_shrimps

Moujahid, A., Inza, I., & Larrañaga, P. (2000). *sc.ehu.es.* Fonte: http://www.sc.ehu.es/ccwbayes/docencia/mmcc/docs/t9knn.pdf

Nabors, M. W. (15 de Julio de 2011). *biblio3.url.edu.gt.* Fonte: http://biblio3.url.edu.gt/Libros/2011/bot/18.pdf

Narro Ramirez, A. E. (1996). *Aplicación de algunos modelos matempáticos para la toma de decisiones.* Fonte: redalyc.org: https://www.redalyc.org/pdf/267/26700614.pdf

Navas Ureñas, J. (2018). *matema.ujaen.es.* Fonte: http://matema.ujaen.es/jnavas/web_modelos_empresa/archivos/archivos%20pdf/teoria/teoria%20discreto/teoria%20discreto%20tema1.pdf

Nguyen Tang, K., Nguyen Dinh, T., Tra Hoang, S., & Luong Hong, D. (Febrero de 2015). *researchgate.net.* Fonte: https://www.researchgate.net/publication/308847347_Automated_monitoring_and_control_system_for_shrimp_farms_based_on_embedded_system_and_wireless_sensor_network

Nicovita. (2018). *nicovita.com.pe.* Fonte: https://www.nicovita.com.pe/SitePages/es/productos.aspx

Pinedo Cortés, L. A. (2017). *la computación en méxico por especialidades académicas.* Mexico: Academia Mexicana de Computación, A. C.

Produce. (2008). *Academia.edu.* Fonte: https://www.academia.edu/7222358/Sistema_productivo_de_Langostinos?email_work_card=view-paper

Quevedo, A., Vegas, S., Loda, J., Cedino, G., & Vining, G. (2020). *Within batch non-linear profile monitoring applied to shrimp farming. Manuscript Submitted for publication*. Blacksburg: Virginia Tech Publishing.

Ríos, S. (1995). *Modelización.* Madrid: Alianza Universidad.

Rivas M., G., López, L. A., & Velasco, A. (2004). Regresión no lineal. *Revista Colombiana de estadística*, 89-102.

Serna M., E., Serna, A., & Acevedo, E. (Septiembre de 2017). *researchgate.net.* Fonte: https://www.researchgate.net/publication/331498946_Principios_y_caracteristicas_de_las_redes_neuronales_artificiales

Sociedad de Comercio Exterior del Perú. (Febrero de 2018). *comexperu.org.pe.* Acesso em 22 de Octubre de 2019, disponível em https://www.comexperu.org.pe/articulo/exportaciones-de-langostinos-alcanzan-record

Ulloa Ibarra, J., & Rodríguez Carrillo, J. (Marzo de 2010). *redalyc.org.* Fonte: https://www.redalyc.org/pdf/636/63613123024.pdf

Ulloa Tello, R. F. (2015). *El efecto de dos porcentajes de recirculación de agua en el cultivo de camarón (Litopenaeus vannamei).* Machala: Universidad Técnica de Machala.

White, R. (1992). *The detection and testing of multivariate outliers.* Vancouver: university of british columbia.

Winsor, C. (1932). *The Gompertz Curve as a Growth curve.* Baltimore: National Academy of Sciences. Fonte: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC1076153/pdf/pnas01729-0009.pdf

Xijun Tian, P., & Hochman, E. (1993). *Shrimp Growth Functions and Their Econom Implications.* Honolulu: Aquacultural Engineering.

Yépez Pinillos, V. (Octubre de 2002). *IMARPE.* Fonte: http://www.imarpe.pe/imarpe/archivos/informes/acuic_esta_sit_maricult.pdf

UNIVERSITAS STUDIORUM PIURENSIS

Anexos

UNIVERSITAS STUDIORUM PIURENSIS

Anexo A. Tabla del resumen estadístico de los pesos por semana.

	Media	Desviación	Variación
2	2.212978	0.7310173	0.5343863
3	2.830345	0.9835199	0.9673113
4	3.691986	1.2101794	1.4645342
5	4.670069	1.4836437	2.2011986
6	5.751301	1.7601206	3.0980244
7	6.890651	2.1637034	4.6816123
8	8.115486	2.4859652	6.1800228
9	9.369167	2.7390638	7.5024704
10	10.565226	2.8104585	7.8986768
11	11.755972	2.8373955	8.0508135
12	13.064239	2.8246297	7.9785329
13	14.311359	2.8509895	8.1281413
14	15.585568	3.0110839	9.0666263
15	16.698099	3.0655594	9.3976544
16	17.858750	3.1434216	9.8810992
17	18.882243	3.1697676	10.0474266
18	19.776809	3.2192643	10.3636624
19	20.615627	3.2784961	10.7485369
20	21.295213	3.2082086	10.2926022
21	21.721667	3.0325371	9.1962811
22	22.361593	2.8390069	8.0599599
23	22.987164	2.7413273	7.5148752
24	24.069355	1.9937752	3.9751396
25	24.752941	1.9006143	3.6123346
26	25.446667	3.2946967	10.8550267
27	28.566667	0.5131601	0.2633333
28	28.900000	NA	NA

Anexo B. Tabla de la media de las variables más importantes de las 100 primeras piscinas.

	P_Nicov	Fact_Conv	T_Max	O2_Max	pH_Min	Alcalinidad	Cianofitas
P(1-01)	0.0559333	1.7177378	27.82333	8.588889	7.800000	93.33333	1.3833333
P(1-02)	1.0000000	1.3267187	28.10619	8.137619	8.381818	119.37500	2.7559783
P(1-03)	1.0000000	1.2583465	27.31591	8.368182	8.454546	124.58333	1.1880682
P(1-04)	1.0000000	1.0933525	27.04715	8.154290	8.450000	141.08330	1.8750000
P(1-05)	1.0000000	1.2052572	27.08056	7.809528	8.450000	131.54757	3.9916667
P(1-06)	0.9583333	1.4033133	27.11569	8.313500	8.354546	131.42850	1.2218750
P(1-07)	0.9600000	1.3459583	27.16682	8.804462	8.536364	111.12500	1.1119565
P(1-08)	1.0000000	1.3737600	27.93613	8.127168	8.200000	131.70455	1.7940217
P(1-09)	0.0000000	1.1321289	27.72792	7.884778	8.250000	129.33337	2.0229167
P(1-11)	1.0000000	1.1441355	26.94794	7.964435	8.350000	141.42857	3.3600000

Anexo C. Tabla de la varianza de las variables más importantes de las 100 primeras piscinas

	P_Nicov	Fact_Conv	T_Max	O2_Max	pH_Min	Alcalinidad	Cianofitas
P(1-01)	0.0164950	0.1943459	0.3980000	0.4936111	NA	8.333333	1.9501667
P(1-02)	0.0000000	0.1100833	1.9942348	1.3885690	0.0836364	1393.229167	12.7632084
P(1-03)	0.0000000	0.1269935	1.9997110	1.5222727	0.0947273	861.041667	0.6898434
P(1-04)	0.0000000	0.1103340	1.9103775	0.9853885	0.0761111	1850.626500	3.6963322
P(1-05)	0.0000000	0.0884240	1.9513855	1.0555119	0.0885714	438.992341	18.4027206
P(1-06)	0.0416667	0.0916803	1.5417888	2.1304009	0.1847273	542.243673	0.7084952
P(1-07)	0.0400000	0.1097405	1.9203549	1.9023661	0.1565455	1332.916667	2.5796091
P(1-08)	0.0000000	0.1450979	1.7234794	1.3252816	0.0260000	778.835227	4.9507936
P(1-09)	0.0000000	0.1139036	2.0821689	0.8121685	0.0457143	1007.839111	3.1496186
P(1-11)	0.0000000	0.0996632	1.6039053	0.9677561	0.1050000	639.285714	15.4081184