Universidad de San Carlos de Guatemala
Facultad de Ingeniería
Escuela de Estudios de Postgrado
Maestría en Artes en Estadística Aplicada

APLICACIÓN DE ANÁLISIS DE REGRESIÓN PARA ESTIMACIÓN DE PRODUCTIVIDAD DE CAÑA DE AZÚCAR UTIILIZANDO EL ÍNDICE DE VEGETACIÓN DE DIFERENCIA NORMALIZADA (NDVI) EN LA ZONA CAÑERA DE GUATEMALA

Ing. Carlos Vinicio Mazariegos Tello
Asesorado por el Mtro. Ing. Joel Estuardo Morales Lemus

Guatemala, febrero de 2022

UNIVERSIDAD DE SAN CARLOS DE GUATEMALA

FACULTAD DE INGENIERÍA

APLICACIÓN DE ANÁLISIS DE REGRESIÓN PARA ESTIMACIÓN DE PRODUCTIVIDAD DE CAÑA DE AZÚCAR UTIILIZANDO EL ÍNDICE DE VEGETACIÓN DE DIFERENCIA NORMALIZADA (NDVI) EN LA ZONA CAÑERA DE GUATEMALA

TRABAJO DE GRADUACIÓN

PRESENTADO A LA JUNTA DIRECTIVA DE LA
FACULTAD DE INGENIERÍA
POR

ING. CARLOS VINICIO MAZARIEGOS TELLO
ASESORADO POR EL MTRO. JOEL ESTUARDO MORALES LEMUS

AL CONFERÍRSELE EL TÍTULO DE

MAESTRO EN ESTADÍSTICA APLICADA

GUATEMALA, FEBRERO DE 2022

UNIVERSIDAD DE SAN CARLOS DE GUATEMALA

FACULTAD DE INGENIERÍA

NÓMINA DE JUNTA DIRECTIVA

DECANA	Inga. Aurelia Anabela Cordova Estrada
VOCAL I	Ing. José Francisco Gómez Rivera
VOCAL II	Ing. Mario Renato Escobedo Martínez
VOCAL III	Ing. José Milton de León Bran
VOCAL IV	Br. Kevin Vladimir Armando Cruz Lorente
VOCAL V	Br. Fernando José Paz González
SECRETARIO	Ing. Hugo Humberto Rivera Pérez

TRIBUNAL QUE PRACTICÓ EL EXAMEN GENERAL DE DEFENSA

DECANA	Inga. Aurelia Anabela Cordova Estrada
DIRECTOR	Mtro. Ing. Edgar Darío Álvarez Cotí
EXAMINADOR	Mtro. Ing. Edwin Bracamonte Orozco
EXAMINADOR	Mtro. Ing. William Eduardo Fagiani Cruz
SECRETARIO	Ing. Hugo Humberto Rivera Pérez

HONORABLE TRIBUNAL EXAMINADOR

En cumplimiento con los preceptos que establece la ley de la Universidad de San Carlos de Guatemala, presento a su consideración mi trabajo de graduación titulado:

APLICACIÓN DE ANÁLISIS DE REGRESIÓN PARA ESTIMACIÓN DE PRODUCTIVIDAD DE CAÑA DE AZÚCAR UTIILIZANDO EL ÍNDICE DE VEGETACIÓN DE DIFERENCIA NORMALIZADA (NDVI) EN LA ZONA CAÑERA DE GUATEMALA

Tema que me fuera asignado por la Dirección de la Escuela de Estudios de Postgrado, con fecha 15 de noviembre de 2020.

Ing. Carlos Vinicio Mazariegos Tello

Decanato
Facultad de Ingeniería
24189101- 24189102
secretariadecanato@ingenieria.usac.edu.gt

LNG.DECANATO.OI.0119.2022

La Decana de la Facultad de Ingeniería de la Universidad de San Carlos de Guatemala, luego de conocer la aprobación por parte del Director de la Escuela de Estudios de Posgrado, al Trabajo de Graduación titulado: **APLICACIÓN DE ANÁLISIS DE REGRESIÓN PARA ESTIMACIÓN DE PRODUCTIVIDAD DE CAÑA DE AZÚCAR UTIILIZANDO EL ÍNDICE DE VEGETACIÓN DE DIFERENCIA NORMALIZADA (NDVI) EN LA ZONA CAÑERA DE GUATEMALA,** presentado por: **Carlos Vinicio Mazariegos Tello ,** que pertenece al programa de Maestría en artes en Estadística aplicada después de haber culminado las revisiones previas bajo la responsabilidad de las instancias correspondientes, autoriza la impresión del mismo.

IMPRÍMASE:

UNIVERSIDAD DE SAN CARLOS DE GUATEMALA
DECANA
FACULTAD DE INGENIERÍA

Inga. Aurelia Anabela Cordova Estrada
Decana

Guatemala, febrero de 2022

AACE/gaoc

Guatemala, febrero de 2022

LNG.EEP.OI.119.2022

En mi calidad de Director de la Escuela de Estudios de Postgrado de la Facultad de Ingeniería de la Universidad de San Carlos de Guatemala, luego de conocer el dictamen del asesor, verificar la aprobación del Coordinador de Maestría y la aprobación del Área de Lingüística al trabajo de graduación titulado:

"APLICACIÓN DE ANÁLISIS DE REGRESIÓN PARA ESTIMACIÓN DE PRODUCTIVIDAD DE CAÑA DE AZÚCAR UTIILIZANDO EL ÍNDICE DE VEGETACIÓN DE DIFERENCIA NORMALIZADA (NDVI) EN LA ZONA CAÑERA DE GUATEMALA"

presentado por **Carlos Vinicio Mazariegos Tello** correspondiente al programa de **Maestría en artes en Estadística aplicada** ; apruebo y autorizo el mismo.

Atentamente,

"Id y Enseñad a Todos"

UNIVERSIDAD DE SAN CARLOS DE GUATEMALA · FACULTAD DE INGENIERÍA · DIRECCIÓN · ESCUELA DE POSTGRADO

Mtro. Ing. Edgar Darío Álvarez Cotí
Director
Escuela de Estudios de Postgrado
Facultad de Ingeniería

Guatemala 23 de septiembre 2021.

M.A. Edgar Darío Álvarez Cotí
Director
Escuela de Estudios de Postgrado
Presente

M.A. Ingeniero Álvarez Cotí:

Por este medio informo que he revisado y aprobado el Informe Final del trabajo de graduación titulado **"APLICACIÓN DE ANÁLISIS DE REGRESIÓN PARA ESTIMACIÓN DE PRODUCTIVIDAD DE CAÑA DE AZÚCAR UTIILIZANDO EL ÍNDICE DE VEGETACIÓN DE DIFERENCIA NORMALIZADA (NDVI) EN LA ZONA CAÑERA DE GUATEMALA"** del estudiante **Carlos Vinicio Mazariegos Tello** quien se identifica con número de carné **201790709** del programa de Maestría en Estadística Aplicada.

Con base en la evaluación realizada hago constar que he evaluado la calidad, validez, pertinencia y coherencia de los resultados obtenidos en el trabajo presentado y según lo establecido en el *Normativo de Tesis y Trabajos de Graduación aprobado por Junta Directiva de la Facultad de Ingeniería Punto Sexto inciso 6.10 del Acta 04-2014 de sesión celebrada el 04 de febrero de 2014*. Por lo cual el trabajo evaluado cuenta con mi aprobación.

Agradeciendo su atención y deseándole éxitos en sus actividades profesionales me suscribo.

Atentamente,

MSc. Ing. Edwin Adalberto Bracamonte Orozco
Coordinador
Maestría en Estadística Aplicada
Escuela de Estudios de Postgrado

Guatemala, 10 de septiembre de 2021.

Ing. Edgar Darío Álvarez Cotí
Director de la Escuela de Estudios de Postgrado. FIUSAC.
Presente.

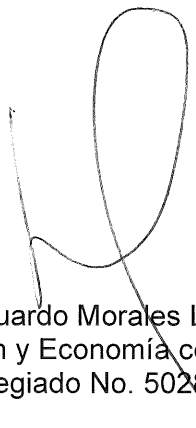
Estimado Ingeniero Álvarez Cotí:

Le saludo esperando que goce de una excelente semana. Por medio de la presente hago de su conocimiento que Carlos Vinicio Mazariegos Tello, estudiante de la Maestría en Estadística Aplicada, quien se identifica con carné número 201790709, me ha presentado el informe final y artículo científico de su Trabajo de Graduación titulado "Aplicación de análisis de regresión para estimación de productividad de caña de azúcar utilizando el índice de vegetación de diferencia normalizada (NDVI) en la zona cañera de Guatemala", solicitando mi revisión y aprobación.

Luego de revisar el documento que contiene el informe final y artículo científico de la investigación, manifiesto que el estudio se ha realizado bajo mi asesoría, de acuerdo con lo establecido en el protocolo aprobado para ello y que los resultados obtenidos, son los esperados. Por lo cual doy por concluido el estudio y le otorgo mi aprobación al informe elaborado.

Sin otro particular, me suscribo a sus respetables órdenes.

Atentamente,



Joel Estuardo Morales Lemus
Maestro en Administración y Economía con Énfasis en Finanzas
Colegiado No. 5028

ACTO QUE DEDICO A:

Dios	Por todas las bendiciones recibidas.
Mis padres	Guillermina Tello y Carlos Mazariegos, su amor será siempre mi inspiración.
Mis hermanos	Alex Gabriel, María Gabriela y Fernando José Mazariegos, por ser una importante influencia en mi carrera.
Mis tíos	Miriam y Silvia Tello y Eric Gordillo, por ser siempre apoyo incondicional.
Mi abuela	Cristina López, por ser una importante influencia en mi carrera, entre otras cosas.
Mi amada esposa	Ilse Barillas, por todo su apoyo.
Mi hija	Emma Mazariegos, para que luche por sus sueños

AGRADECIMIENTOS A:

Universidad de San Carlos de Guatemala	Por permitirme culminar mis estudios de maestría y ayudarme a crecer como profesional.
Mi asesor	Ing. Joel Morales, por ser una importante influencia en mi carrera, y crecimiento profesional.
Mis revisoras	Sc. D. Mayra Castillo, Sc. D. Aura Rodríguez, por apoyarme en esta etapa de estudios.
Mis amigos	Ing. Francisco Pec y Willy Tut, por el acompañamiento brindado.

ÍNDICE GENERAL

ÍNDICE DE ILUSTRACIONES

FIGURAS

TABLAS

LISTA DE SÍMBOLOS

Símbolo	Significado
CV	Coeficiente de variación
s	Desviación estándar
NDVI	Índice vegetativo de diferencia normalizada
kgAzTc	Kilogramos de azúcar por tonelada de caña
m	Metro
mm	Milímetro
nm	Nanómetro
%	Porcentaje
RMSE	Raíz del error cuadrático medio
∑	Sumatoria
TAH	Toneladas de azúcar por hectárea
TCH	Toneladas de caña por hectárea
s^2	Varianza

GLOSARIO

Productividad	Hace referencia a la producción de biomasa.
Rendimiento	Hace referencia a la producción de kilogramos de azúcar por tonelada de caña.
Zafra	Sinónimo de cosecha de caña de azúcar.

RESUMEN

El propósito de la investigación es la aplicación del análisis de regresión para generar un modelo que tome en cuenta la variabilidad espacial para el pronóstico de la productividad de caña de azúcar, como una herramienta de planificación de un ingenio azucarero de Guatemala.

El objetivo general es determinar el modelo de mejor ajuste, utilizando el índice vegetativo de diferencia normalizada (NDVI), que proyecta la producción del ingenio azucarero.

La metodología toma en cuenta el análisis exploratorio de datos, el método de selección secuencia hacia atrás y el análisis de regresión para obtener modelos que representen la variabilidad de la productividad, el mejor modelo se seleccionó en base al criterio de información de Akaike y se validaron los supuestos.

El análisis de correlación lineal de Pearson determinó relación alta positiva para las categorías elongación 1, elongación 2, acuinmacel12 y acuinmael1. El mejor modelo presentó un R^2 de 0.75 y un AIC de 1236, donde las variables significativas del modelo fueron macollamiento, elongación 1 y el acumulado de iniciación, macollamiento y elongación 1. La investigación beneficia al ingenio azucarero con una herramienta de predicción de la productividad, objetivo, con un nivel de certeza y confianza conocido, que incide directamente en la programación de actividades y recursos.

El análisis de varianza determinó que la regresora que más explica la variabilidad de la productividad fue la etapa de elongación 1, en un 41.16 %. Para próximos estudios se recomienda al investigador, evaluar el grado de ajuste del NDVI con imágenes de mayor resolución para mejorar el ajuste del modelo.

PLANTEAMIENTO DEL PROBLEMA

- Contexto general

El problema estadístico identificado se relacionó a la necesidad de un modelo de predicción de la productividad de caña de azúcar, objetivo y preciso, con un nivel de certeza y confianza conocido, que brinde al ingenio mayor exactitud en la programación de actividades y recursos, porque actualmente el estimado se hace en base a la experiencia de los técnicos o con muestreos estadísticos, que no toman en cuenta la variabilidad espacial de la unidad productiva.

La estimación de la producción de caña y sacarosa constituye el punto de partida de la planificación de las actividades de cosecha, transporte y molienda de caña, elaboración del presupuesto de operación, cumplimiento de las entregas de cuotas de azúcar, estimados de ingresos por la venta del azúcar y los subproductos en los ingenios azucareros. La evaluación del potencial de productividad de la caña de azúcar se realiza generalmente al finalizar la cosecha, cuando ya se tiene el registro de la productividad del área, basándose en la experiencia de los técnicos o bien con muestreos estadísticos, sin considerar la variación espacial por diversos factores ambientales, lo que genera errores en la estimación de la producción (Subirós, Sánchez, Esquivel, 2010).

Un desfase alto ocurre especialmente cuando el rendimiento real está por debajo del estimado y tiene costosas repercusiones a nivel operativo, administrativo y financiero. Ante la problemática del nivel de incertidumbre que generan los estimados de campo por métodos subjetivos, basados en la

experiencia del encargado de la finca, se empezó a utilizar en el ingenio el Índice vegetativo de diferencia normalizada(NDVI), como un estimador de la producción, se han utilizado modelos de regresión lineal múltiple, usando como variables independientes el NDVI del año actual, el NDVI de la zafra 2015-2016 y NDVI de la zafra 2014-2015, para predecir la producción de toneladas de caña por hectárea.

Tabla I. **Estimados de productividad de la zafra 2016-2017**

Tercio de zafra	Productividad	Zafra 2016-2017 (TCH)	Variación vs Real (%)
Primero	Estimado de campo	127	9
	Estimado NDVI	122	4
	Real	117	
Segundo	Estimado de campo	110	3
	Estimado NDVI	98	-8
	Real	107	

Fuente: elaboración propia.

En la tabla 1 se muestra la variación que existió en el estimado de producción respecto a la biomasa real, para el primero, el estimado del índice de vegetación de diferencia normalizada fue el más cercano a la producción real (4 % de variación); sin embargo, en el segundo período, el más cercano fue el estimado de campo (3 % de variación). Ambos estimadores de producción presentaron un nivel de incertidumbre parecido. El modelo del ingenio obtuvo un coeficiente de determinación de 0.68 y un error típico de 16.59 toneladas de caña por hectárea, lo cual podría ser objeto de mejora.

- Descripción del problema

El ingenio azucarero situado en la costa sur de Guatemala administra 60,000 hectáreas de producción de caña de azúcar y la determinación del estimado de producción se realizó basado en el historial de producción de la unidad productiva, según la experiencia del técnico o con muestreos estadísticos, pero ninguno consideró la variación espacial debido a diversos factores ambientales, ni el tipo de índice oceánico El Niño (ONI, por sus siglas en inglés) predominante durante el año. El problema estadístico identificado se enfocó a la falta de un modelo de predicción de la productividad, objetivo y preciso, con un nivel de certeza y un nivel de confianza conocido, para predecir la productividad de caña de azúcar del ingenio azucarero. Esta carencia impidió precisar mejores resultados en oportunidad, detalle y calidad de información que permitiera al ingenio mejorar la eficiencia de sus procesos a lo largo de la cadena de valor, definir con mayor certeza el recurso para el proceso de cosecha, producción de azúcar en la fábrica, obtener mayor certeza en proyecciones de ventas, recursos del siguiente ciclo productivo y presupuestos.

- Formulación del problema

    - Pregunta central

        - ¿Cuál es el modelo de mejor ajuste, utilizando el Índice vegetativo de diferencia normalizada (NDVI), que proyecta la producción del ingenio azucarero por tercio de zafra?

- Preguntas auxiliares

    - ¿Cuáles son las categorías potenciales de NDVI para la predicción de productividad de caña de azúcar?

    - ¿Cuáles son las variables regresoras (categorías) que contribuyen a obtener un mayor grado de ajuste del modelo?

    - ¿Cuál es el grado de ajuste del modelo que mejor representa la relación entre la productividad y la o las variables regresoras?

    - ¿El modelo de mejor ajuste cumple con los principales supuestos del análisis de regresión?

- Delimitación del problema

El estudio se realizó con la información generada durante la zafra 2018-2019, que inició desde noviembre de 2017 y finalizó en abril de 2019. Se recolectó el valor del NDVI desde noviembre de 2017 hasta abril de 2019, y se obtuvieron los registros de productividad que corresponden al mismo ciclo de producción, del perodo de noviembre de 2018 a abril de 2019, que abarcaron 3000 hectáreas en total, representando el 5 % del total de área administrada por el ingenio azucarero.

OBJETIVOS

General

Determinar el modelo de mejor ajuste, utilizando el índice vegetativo de diferencia normalizada (NDVI), que proyecta la producción del ingenio azucarero.

Específicos

- Establecer las categorías potenciales de NDVI para la predicción de productividad de caña de azúcar.

- Seleccionar las variables regresoras (categorías) que contribuyan a obtener un mayor grado de ajuste del modelo.

- Determinar el grado de ajuste del modelo que mejor representa la relación entre la productividad y la o las variables regresoras.

RESUMEN DEL MARCO METODOLÓGICO

El enfoque del trabajo de investigación fue cuantitativo con variables continuas, donde a partir de análisis de Índices espaciales, se determinaron modelos de predicción de productividad de caña de azúcar para un ingenio azucarero.

El diseño de la investigación no fue experimental, porque no se realizaron experimentos para determinar el resultado de las variables dependientes, es decir, no se alteraron variables de manera intencional para analizar los resultados. Los datos se obtuvieron de imágenes satelitales que generaron información del índice vegetativo de diferencia normalizada (NDVI) y se analizaron en su estado natural, sin manipulación previa.

Se seleccionó un estudio de pronósticos mediante regresión lineal, porque se pretendía generar un modelo de mejor ajuste, para proyectar la productividad de caña de azúcar en toneladas de caña por hectárea, por tercio de cosecha del ingenio azucarero.

Tuvo un alcance descriptivo y correlacional, porque se analizaron bases de datos del índice vegetativo de diferencia normalizada y base de productividad de un ingenio azucarero, para determinar un modelo de certeza y confianza conocida, que permitió mejorar el estimado de producción de caña de azúcar.

Las variables del estudio fueron:

- Índice vegetativo de diferencia normalizada de las unidades productivas del ingenio azucarero, obtenido por medio de imágenes del satélite *Sentinel,* a lo largo del ciclo productivo de caña de azúcar de la temporada 2018-2019.

- Coeficiente de determinación por método de eliminación secuencial hacia atrás (*backward elimination,* en inglés): determinado por medio de la combinación de variables que aportaron significativamente al modelo.

- Criterio de información de Akaike: determinado por el estadístico de Akaike que permitió escoger el mejor modelo que mejoró el ajuste y redujo la pérdida de información.

- Prueba de normalidad de los residuos por Kolmogorov-Smirnov, independencia por Durbin-Watson, heterocedasticidad por Breusch-Pagan. Las variables e indicadores del estudio se describen a continuación.

Tabla II. **Variables e indicadores**

Variable	Definición	Indicador
Índice de diferencia normalizada	Valor representativo de índice vegetativo de diferencia normalizada (NDVI) presente el en momento del ciclo de producción.	Distribución simétrica de las observaciones
Índice de Akaike (AIC)	Estimación de la distancia entre el modelo y el mecanismo que genera los datos observados, determina el modelo que pierde la menor cantidad de información.	Modelo de regresión con menor criterio de información de Akaike

Continuación tabla II.

Coeficiente de determinación del método eliminación hacia atrás.	Método que pretende reducir las variables necesarias para el modelo y cumplir con el principio de parsimonia, a través del mejor coeficiente de determinación.	Modelos con mayor coeficiente de determinación
Prueba de normalidad Kolmogorov-Smirnoff, independencia Durbin-Watson, heterocedasticidad Breush-Pagan	Pruebas estadísticas que validan los supuestos de la regresión.	Validar que el modelo cumple la hipótesis de normalidad, homocedasticidad e independencia

Fuente: elaboración propia.

- Fases de la investigación

    - Fase 1: revisión de literatura

Para fundamentar la investigación y planteamiento de la propuesta del uso de series de tiempo y pronósticos para estimar la productividad de caña de azúcar, se recopiló información de libros, artículos y trabajos realizados con respecto al tema de uso del índice de diferencia normalizada, series temporales y pronósticos, trabajos realizados en estimación de rendimiento de caña de azúcar y temas relacionados al uso de imágenes satelitales en predicción de rendimiento.

El resultado obtenido de esta fase fue establecer la base teórica como marco de referencia para el desarrollo del trabajo de investigación.

- Fase 2: gestión de la información

Por medio del software de cálculo Excel®, se procedió a depurar y categorizar la base de datos del Índice de vegetación de diferencia normalizada proveniente de la red de nano satélites *Planet*® y Satélite *LandSat* y su respectivo valor de productividad (toneladas de caña por hectárea), correspondientes al año 2017 y 2018, cubriendo alrededor de 50,000 hectáreas de área productiva de caña de azúcar de la costa sur de Guatemala. Se realizó la categorización de cada valor del índice vegetativo de diferencia normalizada.

- Fase 3: análisis de la información

Para el análisis descriptivo se utilizó el paquete estadístico Infostat® versión 2008, en el cual se ingresó la base de datos de NDVI y productividad, proveniente del paquete de cálculo Excel®.

El método de eliminación secuencial hacia atrás (*backward elimination*) se realizó con el paquete estadístico Infostat®, donde se seleccionaron las variables que aportan al modelo.

Las pruebas estadísticas criterio de información de Akaike (AIC), Breusch-Pagan, Kolmogorov-Sminov, Durbin-Watson, se realizaron en el paquete estadístico R®

- Fase 4: interpretación de la información

El método de selección secuencial hacia atrás permitió identificar las variables que aportan significativamente al modelo, por medio del análisis de regresión se calcularon los modelos de mejor ajuste, aquellos que presentaron el

mayor coeficiente de determinación. Por medio del criterio de información de Akaike, se seleccionó el modelo de mejor ajuste entre los modelos generados, por medio de las pruebas de normalidad, independencia y heterocedasticidad, se validó el modelo que cumplió con los supuestos del análisis de regresión.

- Fase 5: elaboración del informe final

Se redactaron los resultados obtenidos, compilando toda la información analizada en un documento.

- Población y muestra

La población del estudio fueron imágenes satelitales correspondiente a 5000 hectáreas cultivadas con caña de azúcar, ubicadas en la zona cañera de la costa sur de Guatemala, pertenecientes a un ingenio azucarero del año 2017 y 2018.

- Técnicas de análisis de la información

Se describen las técnicas estadísticas para el análisis de la información.

- Análisis exploratorio de datos

Los datos obtenidos de las imágenes satelitales se tabularon y ordenaron en una hoja de cálculo Excel®, colocando de forma ordenada en columnas todas las variables para el análisis. Por medio de estadística descriptiva se obtuvieron las medidas de tendencia central, medidas de dispersión, se determinó la existencia de datos atípicos. Se generaron categorías en la serie de tiempo de NDVI que se usaron como variables regresoras en el cálculo de los modelos de regresión maximizando el R^2 y minimizando el AIC.

- Selección secuencial hacia atrás (*backward elimination*)

Se categorizó la serie de datos del NDVI, según las categorías descritas que aportan al coeficiente de determinación (R^2), para cumplir el principio de parsimonia y evitar colinealidad en los modelos generados.

- Generación de modelos por análisis de regresión

Con el paquete estadístico R® se generaron los modelos de regresión, usando como variables predictivas las categorías significativas del índice vegetativo de diferencia normalizada y como variable dependiente la productividad en toneladas de caña por hectárea, se seleccionaron los modelos con mejor ajuste, que presentaron los mayores valores de coeficiente de determinación (R^2).

- Selección del mejor modelo por criterio de información de Akaike

Se seleccionaron los modelos que presentaron el menor criterio de información de Akaike, para asegurar que los modelos seleccionados presenten el mejor ajuste y la menor pérdida de información.

- Validación de supuestos del análisis de regresión

Se validaron los supuestos de la regresión en el modelo de mejor ajuste, por medio de la prueba de normalidad de residuos aplicando la prueba Kolmogorov-Smirnov, independencia por medio de la prueba Durbin-Watson y heterocedasticidad por medio de Breush-Pagan.

- Validación del modelo de mejor ajuste

La validación se realizó comparando el pronóstico con los valores observados en cada período estacional, usando 100 datos de la temporada 2017 que no se usaron en la generación del modelo, mediante un análisis de correlación de Pearson, se midió el grado de asociación entre los valores estimados y los valores reales.

INTRODUCCIÓN

El trabajo se enfoca en la aplicación del análisis de regresión para el desarrollo de un modelo objetivo y preciso que pronostique la productividad de caña de azúcar, por medio del estudio de la variabilidad espacial, dada por la variable de índice de vegetación de diferencia normalizada (NDVI). Es un trabajo de sistematización porque pretende cambiar la forma tradicional de realizar el pronóstico de producción de caña de azúcar, el cual está basado en la experiencia del administrador de la finca, en el historial productivo o en muestreos estadísticos, pero ninguno toma en cuenta la variación espacial debido a factores ambientales (Subirós, Sánchez y Esquivel, 2010).

El problema radica en que el pronóstico de producción es la base para la planificación de actividades, elaboración de presupuestos, proyecciones de ingresos por la venta de azúcar, cumplimiento de entregas, generación de contratos de maquinaria agrícola y cosecha, entre otros. Ante la falta de certeza y errores de estimación que genera un pronóstico empírico, se producen desperdicios en toda la cadena de producción de azúcar.

La importancia del estudio es generar modelos que tomen en cuenta la variabilidad espacial para pronósticos de corto plazo de la productividad de caña de azúcar, para que los departamentos asociados a la cadena de valor, tales como Cosecha, Finanzas, Campo y Fábrica, puedan ajustar sus recursos y procesos a un estimado de producción de menor incertidumbre y mayor precisión y mejorar los presupuestos parciales.

Para probar el punto de la evaluación se utiliza estadística descriptiva, para realizar un análisis exploratorio de la media, moda, mediana, varianza y desviación estándar, para identificar y describir la distribución de las series.

Posteriormente se realiza la prueba de normalidad Kolmogorov-Smirnov para determinar la distribución de las series. Después de corroborar la normalidad se validan los supuestos, mediante análisis de dependencia y homogeneidad de varianzas. Se realiza la categorización de la base de datos del NDVI y productividad de caña, para identificar las categorías con influencia significativa sobre la productividad, con el método de selección secuencial hacia atrás (*Backward elimination*,), con un nivel de significancia de 0.10, para eliminar características redundantes en la serie del NDVI. Luego se selecciona el modelo de pronóstico que presente los mejores ajustes para la productividad de caña de azúcar, utilizando el índice de información de Akaike (AIC).

La validación del modelo seleccionado se realiza mediante análisis de correlación de Pearson para determinar el grado de asociación entre los resultados pronosticados y los predichos, análisis de los residuos y coeficiente de error lineal. Posteriormente se validan los supuestos del modelo de regresión, por medio de análisis de dependencia y homogeneidad de varianzas.

La presente investigación es original y sienta un precedente para el ingenio azucarero, pues aporta beneficios que se reflejan en una mejora significativa en el pronóstico de la producción que permita al ingenio, reducir los desfases de presupuestos de todas las áreas que dependen del volumen de producción para definir su plan de actividades.

1. MARCO REFERENCIAL

Los pronósticos de productividad de caña de azúcar en etapas tempranas del cultivo son muy importantes para la operación de un ingenio azucarero, porque definen las estrategias de comercialización del azúcar, ya que, conociendo la capacidad productiva del cultivo, el ingenio define las estrategias de compra y venta para mejorar su competitividad.

Los pronósticos de productividad van desde la observación, basada en la experiencia del productor, métodos estadísticos y hasta el uso de sensores remotos orbitales, como los satélites *LandSat.* La agroindustria azucarera está en búsqueda constante de herramientas que le permitan ser más asertivos en los pronósticos de producción, porque es muy variable entre cada temporada, debido principalmente a las condiciones climáticas predominantes y la heterogeneidad de las condiciones de suelos y microclimas en los sistemas productivos, por lo que la estimación se convierte en una tarea complicada.

Por tal razón, el uso de sensores que generan información espectral, obtenida de satélites, ha demostrado ser una herramienta poderosa frente a estimaciones que engloban las variaciones del clima y de condiciones de suelo.

.

Bastidas y Carbonell (2006), demostraron que el NDVI presenta una relación directa con la productividad, obteniendo modelos con un R^2 de 0.70, los autores relacionaron la integral de un ciclo completo de caña de azúcar y obtuvieron una relación directamente proporcional con la productividad, de este estudio se obtuvo la metodología para estimar la integral

del ciclo productivo de caña de azúcar y realizar correlaciones, para determinar relación entre la productividad del ingenio y el NDVI.

1.1. Estudios previos

La estrategia de gestión agrícola basada en agricultura de precisión ha permitido la creación de índices de vegetación, mediante los sistemas de información geográfica y tecnología de teledetección, es posible calcular índices en cultivos productivos, para estimar el contenido de agua, diversos tipos de estreses, focos de enfermedades, entre otros (Aguilar, Contreras, Galindo y Fortanelli, 2010).

El más utilizado en la agricultura es el Índice de vegetación diferencial normalizado (NDVI en inglés). Considerado el mejor indicador de la biomasa en caña de azúcar y la fuente más utilizada para calcularlo son las imágenes del satélite *LandSat* 8 (CENGICAÑA, 2016).

El estudio anterior permitió fundamentar el uso del Índice de vegetación diferencial normalizado como variable predictora en modelos de regresión para pronósticos de productividad de caña de azúcar. El ingenio azucarero empezó a utilizar el índice de vegetación de diferencia normalizada desde el año 2012, procesando las imágenes satelitales de *LandSat* 8. Para realizar este estudio, se cuenta con información geográfica de cada unidad productiva de NDVI y productividad de caña de azúcar de los años 2012 a 2018.

Aguilar *et al.* (2010), compararon los valores de NDVI con las condiciones reales del cultivo, utilizando imágenes del satélite *Landsat*-7/ETM, obtuvieron un rango de NDVI de -0.3 a 0.3, que coincidió en campo con valores más bajos y presencia de estrés hídrico y valores más altos y mayor vigor del cultivo. Los

autores encontraron una correlación entre un valor bajo de NDVI y presencia de ciertos estreses en el cultivo de caña de azúcar, que podrían afectar negativamente la producción de biomasa; o un valor mayor de NDVI, relacionado a mayor producción de biomasa, debido a mayor vigorosidad del cultivo.

Este estudio se utilizó para fundamentar el uso del índice de vegetación de diferencia normalizada como el mejor indicador de la productividad de la caña de azúcar, con base en que valores superiores de NDVI están asociados con mayor acumulación de biomasa.

Los estudios realizados para predecir la productividad de caña de azúcar usando el NDVI, no han generado modelos de alta certeza CENGICAÑA (2016). El primer estudio, utilizando sensores para determinar la condición del cultivo, se realizó en Brasil; Rudorff y Batista (1990), calcularon un modelo basado en el Índice Relative Vigor Index (RVI, por sus siglas en inglés), obtenido de imágenes del satélite *Landsat* MSS, que incluía variables meteorológicas, tratando de predecir la productividad de caña de azúcar.

El modelo explica un coeficiente de determinación de 69 %, es decir, el modelo explicó en un 69 % la productividad de la caña de azúcar y arroja un error estimado de 10.5 toneladas de caña por hectárea. Estos estudios incluyen el uso de imágenes satelitales de *LandSat* y variables climáticas en modelos de predicción de productividad de la caña de azúcar, este resultado es fundamental porque en este pronóstico se utilizaron imágenes satelitales de *LandSat* 8, nanosatelites *Planet* y variables climáticas para pronosticar la productividad de caña de azúcar.

Otro estudio que realizaron Schmidt, Narciso, Frost y Gers (2000), en el que usaron imágenes del satélite radiómetro avanzado de muy alta resolución

(AVHRR, por sus siglas en inglés), evidenciaron que es posible obtener modelos de producción a través imágenes espaciales estandarizadas, pero los investigadores calcularon un error estándar de 11 toneladas de caña por hectárea. Este estudio evidencia el uso de pronósticos de series temporales para pronosticar la producción de caña de azúcar.

Para reducir el error, en este estudio, se utilizaron imágenes de alta resolución provenientes de nanosatélites *Planet®,* con mayor resolución que los satélites tradicionales. Se utilizó este estudio para organizar de forma similar el análisis de estadística descriptiva de la base de datos del índice de vegetación de diferencia normalizada y realizar las transformaciones logarítmicas (si lo amerita) para obtener una distribución Normal de la base de datos.

Estudios experimentales a nivel de parcelas han demostrado que es posible obtener una fuerte correlación (R>0.9), entre el comportamiento del Índice de vegetación de diferencia normalizada y la edad del cultivo de caña de azúcar, donde se han obtenido ajustes de hasta 93 %; sugiriendo que existe alta correlación entre el estimado de productividad y la producción comercial real (CENGICAÑA, 2016). En el presente estudio, para incrementar la certeza del modelo se realizó la categorización de la base de datos del NDVI, para que cada valor mensual del índice corresponda a la edad del cultivo de caña y su productividad, posteriormente se utilizó el método de selección secuencial hacia atrás (*backward elimination*), para descartar categorías que no aporten significativamente al modelo.

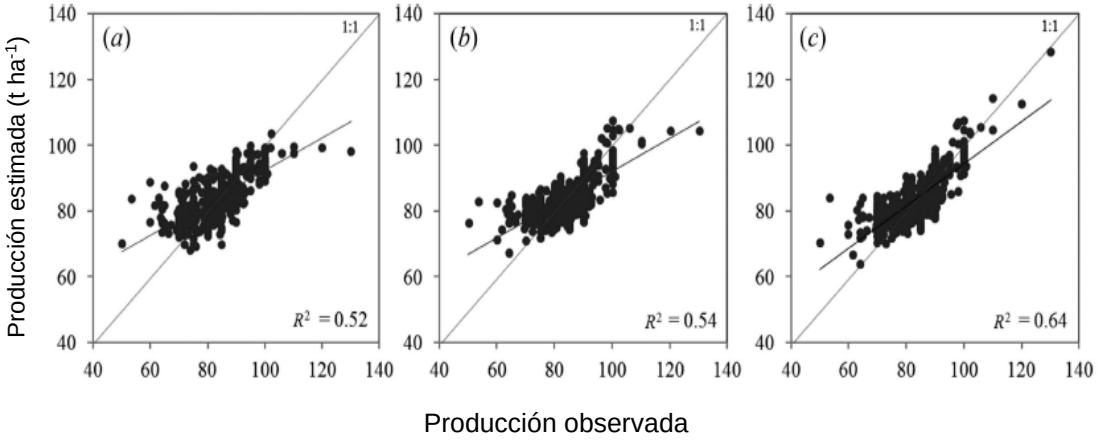
Simões, Rocha y Lamparelli (2005) modelaron la productividad de caña de azúcar, con información generada con sensores remotos de espectrometría, tratando de simular la banda del satélite *Landsat* ETM y obtener valores de NDVI, el modelo generado demostró la alta correlación (0.90) entre la biomasa de caña

de azúcar y el NDVI, obteniendo modelos de regresión múltiple con coeficiente de determinación mayor al 95 %. Los autores mencionados pudieron tener alta certeza porque los sensores incrementaron la resolución del valor NDVI, es decir, un valor para un área más pequeña.

Se utilizó este estudio para determinar el tamaño de la unidad representativa de cada valor del NDVI, es decir, determinar el tamaño de píxel que mejor represente los cambios de la productividad de caña de azúcar.

Los investigadores obtuvieron series temporales de NDVI diferenciadas según el ciclo de cultivo y las sometieron a un modelo de autoaprendizaje de Redes Neuronales Artificiales, para eliminar características redundantes o irrelevantes.

Figura 1. **Relación entre los valores observados y pronosticados con el modelo de regresión lineal múltiple con la data total**



Fuente: Fernandes, Favilla y Dalla. *Sugarcane yield prediction in Brazil using NDVI time series and neural networks ensemble*. Consultado el 20 de enero de 2020. *Recuperado de https://www.researchgate.net/publication/317155546_Sugarcane_yield_prediction_in_Brazil_using_NDVI_time_series_and_neural_networks_ensemble.*

Al someter la data al procedimiento de depuración lograron mejorar el modelo de regresión, porque redujeron la desviación estándar de 8.2 toneladas de caña por hectárea a 5.7 T ha^{-1} e incrementaron el coeficiente de determinación de 0.52 a 0.64. Este estudio se utilizó de base para realizar la categorización de la serie de datos del índice de vegetación de diferencia normalizada, y posteriormente utilizar el método de selección secuencial hacia atrás (*backward elimination*), para eliminar las categorías no significativas de la serie y reducir la desviación estándar de los modelos que se generaron.

Rahman y Robson (2016), utilizaron el método de selección secuencial hacía atrás y determinaron que la variable crítica para estimación de producción de caña de azúcar fue el valor máximo durante todo el ciclo de cultivo del índice vegetativo de diferencia normalizada, utilizando métodos de regresión lineal determinaron un modelo de ajuste significativo (R^2=0.69) entre el valor máximo del NDVI y la productividad de caña de azúcar.

Para validad el modelo utilizaron el método de la desviación de raíz cuadrática media (RMSE), el cual mide las diferencias entre los valores predichos por el modelo y los valores observados, determinaron un RMSE= 4.2 toneladas por hectárea. Los autores mencionados concluyeron que el modelo generado a partir de los valores más altos del NDVI, obtenido de las imágenes del satélite *LandSat,* es una técnica factible para predecir el rendimiento de caña de azúcar en la región de Australia. En este estudio se utilizó el método de desviación de raíz cuadrática media para validar el funcionamiento del modelo de mejor ajuste y determinar la certeza del modelo.

2. MARCO TEÓRICO

2.1. Fundamentos estadísticos

Se realizó una revisión de todos los conceptos asociados al análisis de regresión y los estadísticos que permitieron explicar el comportamiento de cada modelo, con el fin de contar con un marco para la selección del mejor modelo de regresión.

2.1.1. Estadística descriptiva

La estadística descriptiva analiza los grupos de datos, de tal forma que sus métodos de exploración permitan entender su estructura, describiendo las características de mayor importancia del grupo de datos, principalmente para caracterizar las variables, en cuanto a su distribución, centralización y dispersión, se representan a través de medidas de tendencia central, dispersión y gráficos (Aroca, García y González, 2015).

2.1.2. Media

Es una medida de tendencia central, que mide el valor medio de un grupo de datos, tiende a ser afectado por los valores extremos. “La media aritmética se usa como índice de centralización en muestras grandes y variables que siguen una distribución normal y es con mucho la más utilizada” (Aroca *et al.,* 2015, p. 428).

$$\bar{X} = \frac{\sum_i^n x_i}{n} \qquad (Ec. 1)$$

Donde n es el tamaño de la muestra.

2.1.3. Moda

Es una medida de tendencia central que mide el valor de mayor frecuencia en una serie de datos o grupo. "La moda es el valor más repetido de la distribución. Una distribución normal es unimodal (esto es, hay una única moda, que coincide con la media y la mediana)" (Aroca *et al*., 2015).

2.1.4. Mediana

Es el valor de la posición media de una serie de datos o grupo de datos ordenados de forma ascendente o viceversa. "Es el valor que divide a la muestra en dos partes iguales, una vez ordenadas todas las medidas de menos a mayor" (Aroca *et al*., 2015).

2.1.5. Varianza

Es una medida de dispersión que representa el valor medio de la diferencia cuadrática del valor observado respecto a la media del grupo de datos. Para definir la dispersión de las observaciones se calcula la diferencia entre la media aritmética y cada observación, pero genera el problema de que la mitad de los datos serán negativos y el resultado siempre será cero, para eliminar los valores negativos se eleva cada diferencia al cuadrado y se promedian, pero los valores de dispersión son magnificados al cuadrado. (Aroca *et al.,* 2015).

$$s^2 = \frac{\sum_i^n (x_i - \bar{x})^2}{n-1} \qquad (Ec.2)$$

Donde n es el tamaño de la muestra.

2.1.6. Desviación estándar

Pertenece al grupo de medidas de dispersión, se utiliza para obtener la dispersión en las mismas unidades que la media aritmética, al obtener la raíz cuadrada de la varianza (Aroca *et al*., 2015).

$$s = \sqrt{\frac{\sum_i^n (x_i - \bar{x})^2}{n-1}} \qquad (Ec.3)$$

2.1.7. Coeficiente de variación

Mide la dispersión en forma relativa, permite una interpretación más objetiva de la variabilidad, con el coeficiente de variación es posible establecer rangos que determinen niveles de variabilidad poblacional de homogeneidad o heterogeneidad. "Mide la variación porcentual de los datos respecto a su media" (Aroca *et al*., 2015, p. 449).

$$\%C.V. = \frac{s}{\bar{x}} * 100 \quad (Ec.4)$$

Donde s es la desviación estándar de la muestra y $\bar{x}$ es la media de la muestra.

2.1.8. Histogramas

Este tipo de gráfico es el más utilizado para variables cuantitativas discretas o continuas, se utiliza para describir la distribución de la frecuencia de un conjunto de datos (Aroca *et al.,*2015).

2.1.9. Diagrama de cajas (*boxplot*)

Gráfico comúnmente usado que provee un resumen de variables numéricas, la línea que divide la caja en 2 partes representa la mediana de los datos, los extremos de la caja indican los cuartiles superiores e inferiores, las líneas extremas muestran los valores más altos y bajos sin incluir los datos fuera de tipo (*outlier*, en inglés). (Aroca *et al*., 2015).

2.1.10. Coeficiente de correlación lineal de Pearson

El coeficiente de correlación linear de Pearson se aplica en variables cuantitativas, es un índice que mide el grado de asociación entre distintas variables relacionadas en forma lineal. Sus valores van desde -1 a 1.

Se interpreta que la correlación o grado de asociación entre dos variables X y Y es perfecta positiva, cuando exactamente en la medida que incrementa una de las variables, la otra también aumenta, se representa con valores positivos (Aroca *et al*., 2015). Es perfecta negativa, cuando en la medida que una aumenta, la otra disminuye, se representa con valores negativos. El coeficiente de correlación lineal de Pearson se representa con la letra griega Rho y se define por la siguiente ecuación:

$$r_{xy} = \frac{\frac{\sum XY}{N} - \bar{X}\bar{Y}}{S_x S_y} \qquad (Ec. 5)$$

Donde el coeficiente de correlación de Pearson (r), hace referencia al producto de las sumatorias de las variables X y Y, dividido por el tamaño de observaciones, menos el producto de la media de X y Y, este resultado, dividido entre el producto de las varianzas de X y Y.

2.1.11. Modelos estadísticos

Los modelos estadísticos se basan en el análisis y procesamiento de los datos históricos observados y las correlaciones estadísticas entre ellos, de diversas formas y metodología, a partir de ciertas variables como por ejemplo radiación solar o temperatura (predictores), se puede estimar el valor de variables como humedad relativa, amplitud térmica (predictandos) (Martínez, Rivadeneira y Nieto, 2011).

Los modelos estadísticos para la predicción de productividad de caña de azúcar tienen diferentes grados de complejidad; sin embargo, entre los métodos utilizados están: análogos, métodos de regresión, análisis de correlación canónica, redes neuronales, entre otros. El objetivo es detectar patrones coherentes de comportamiento de la variable de interés. Las características dependerán en parte del número de puntos de observación, el número de variables NDVI y variables climáticas presentes en la región de análisis. Las aproximaciones estadísticas permiten realizar análisis de campo promediados, correlaciones de punto a punto, mientras que con técnicas avanzadas permite analizar oscilaciones en fase y fuera de ella (Martínez *et al*., 2011).

Para realizar pronósticos de productividad con modelos estadísticos, es necesario evaluar la normalidad de las variables y dependencia con el fin de seleccionar la metodología adecuada para el grupo de datos. La información de la variable dependiente se basa en observaciones reales y no incorporan errores debido al modelo numérico. Los modelos estadísticos se consideran buenos cuando la variabilidad observada está dominada por una única fuente de predictibilidad y se sustentan en la condición de estacionalidad del clima (Martínez *et al*., 2011).

2.1.12. Análisis de regresión

Permite construir un modelo matemático en el que relaciona el efecto de una o más variables independientes (X1, X2, X3…Xn) sobre la variable dependiente (Y). Esta técnica, trata de buscar el mejor modelo matemático que relacione el efecto de la variable dependiente (productividad de caña de azúcar) con los valores de las variables independientes (NDVI, variables climáticas) (Jiménez, 2004).

$$y = \beta_0 + \beta_1 * x_1 + \beta_2 * x_2 + \beta_3 * x_3 + \cdots + \beta_n * x_n + \varepsilon \quad (Ec.6)$$

Donde: Y= es la variable dependiente (Productividad de caña de azúcar del ingenio azucarero). X= variables independientes (NDVI, variables climáticas). β_i=es el efecto de la variable independiente sobre la dependiente. β_o = es el efecto medio de las variables independientes sobre la dependiente. ε = valor del residuo.

2.1.13. Comparación de modelos mediante el criterio de información de Akaike

El criterio de información de Akaike (*An Information Criterion*, *AIC*), proporciona un método simple y objetivo que selecciona el modelo más adecuado para caracterizar los datos bajo estudio.

> El objetivo ideal de la selección de modelos es conseguir una traslación perfecta, uno a uno, de manera que no se pierda información durante el proceso de generación del modelo. Este objetivo es imposible porque siempre existirá un número finito de datos que contienen una cantidad limitada de información. (Martínez *et al*, 2009, p. 89)

Por lo tanto, el objetivo real es obtener el modelo que mejor se ajusta a los datos, esto es, el modelo que pierda la menor cantidad de información posible. Este criterio se enmarca en el campo de la teoría de la información, se define como:

$$AIC = -2\log\big(\mathcal{L}(\theta)\big) + 2K \quad (Ec.7)$$

Donde $\log(\mathcal{L}(\theta))$ es el logaritmo de la máxima verosimilitud, que permite determinar los valores de los parámetros libres de un modelo estadístico y K es el número de parámetros libres del modelo.

Esta expresión proporciona una estimación de la distancia entre el modelo y el mecanismo que realmente genera los datos observados, que es desconocido y en algunos casos, imposible de caracterizar. Se selecciona el modelo que presente el menor criterio de información de Akaike (Martínez *et al*, 2009).

2.1.14. Validación de supuestos del modelo

La validez del modelo de regresión y la validez de los estadísticos obtenidos dependen de lo razonable de las simplificaciones asociadas o los supuestos del modelo. La importancia de realizar procedimientos contundentes para validar los supuestos radica fundamentalmente en que ellos inciden en las cualidades de los estimadores de mínimos cuadrados (Behar, 2003). Los supuestos son normalidad de los residuos, homocedasticidad de varianzas, autocorrelación en independencia.

2.1.15. Prueba de normalidad de los residuos

Siegel (1956) indica que, al evaluar como hipótesis inicial que los residuos poseen una distribución normal, es necesario una prueba estadística antes de proseguir con cualquier otra técnica. Para corroborar dicho supuesto existe la prueba Kolmogorov-Smirnov.

La prueba de Kolmogorov-Smirnov se basa en calcular las diferencias entre las frecuencias relativas acumuladas de una distribución teórica dada, donde se toma como hipótesis nula que la distribución se comparará con la normal y son iguales, y las frecuencias relativas acumuladas de la muestra se pueden considerar normales, consiste en realizar la siguiente prueba de hipótesis:

- Hipótesis nula: los residuos analizados siguen una distribución normal.
- Hipótesis alterna: los residuos analizados no siguen una distribución normal.

Para corroborar la hipótesis se deben estandarizar las variables (es decir, restarles la media muestral y dividir entre la desviación estándar a cada dato) y a

partir de esta transformación comparar dichos datos con la normal estandarizada (media = 0, y varianza = 1). Luego calcular las diferencias para cada dato con su par “normal”, entonces el estadígrafo de prueba de la prueba Kolmogorov-Smirnoff es:

$$D_c = Max[F_n(x) - F(x)] \qquad (Ec. 8)$$

Donde: D_c es la mayor diferencia observada, $F_n(x)$ es frecuencia acumulada observada y F(x) es la frecuencia teórica.

Si los valores observados son similares a los observados, el valor Dc será pequeño, cuanto mayor sea la diferencia entre la distribución evaluada y la distribución teórica, mayor será el valor D_c. Por lo tanto, el criterio para la toma de decisión de la hipótesis es el siguiente:

- Si $D_c \leq D_\alpha \Rightarrow$ aceptar la hipótesis nula.
- Si $D_c \geq D_\alpha \Rightarrow$ rechazar la hipótesis nula.

Donde el valor D_α se refiere al valor teórico, se encuentra tabulado en tablas y depende del nivel de confianza y el tipo de distribución que se requiere evaluar.

2.1.16. Heterocedasticidad por Breusch-Pagan

La prueba de Breusch-Pagan se utiliza para probar la heterocedasticidad en un modelo de regresión lineal. Prueba si la varianza de los errores de una regresión depende de los valores de las variables independientes, en ese caso, la heterocedasticidad está presente (Garcia y Ortíz 2017). La prueba de Breusch-Pagan para heterocedasticidad es una prueba χ^2, donde el estadístico de prueba es $n\chi^2$ con k grados de libertad. Prueba la hipótesis nula de homocedasticidad.

Si el valor de la χ^2 es significativo con un valor de p por debajo de un umbral apropiado, entonces la hipótesis nula de homocedasticidad es rechazada y se asume la heteroscedasticidad. Si la prueba de Breusch-Pagan demuestra que hay heterocedasticidad condicional, la regresión original puede ser corregida usando el método de Hansen, utilizando errores estándar robustos o reajustando la ecuación de regresión cambiando o transformando variables independientes.
El supuesto de varianza constante se puede examinar haciendo una regresión de los residuos al cuadrado respecto de las variables independientes, usando la siguiente ecuación:

$$\hat{u}^2 = \gamma_0 + \gamma_1 x + v \qquad (Ec. 9)$$

Donde: $\hat{u}^2$ son los residuos al cuadrado y el número de variables independientes $\gamma \ y \ v$.

Primero se deben calcular los residuales, para aplicar mínimos cuadrados ordinarios al modelo:

$$y = X\beta + \varepsilon \quad (Ec. 10)$$

Donde *y* y $\boldsymbol{\varepsilon}$ son vectores de una matriz $n\chi$, X es una matriz de regresores $n \times p$ (matriz de diseño), el coeficiente $\boldsymbol{\beta}$ indica el intercepto.

Posteriormente, se utiliza la ecuación 11, para calcular la regresión auxiliar, el estadístico de prueba es el resultado del coeficiente de determinación de la regresión auxiliar:

$$LM = nR^2 \quad (Ec. 11)$$

El estadístico de prueba se distribuye asintóticamente como ${X^2}_{p-1}$ bajo la hipótesis nula de homocedasticidad.

2.1.17. Independencia por Durbin-Watson

La independencia de los residuos es uno de los supuestos básicos de los modelos de regresión. "El estadístico de Durbin-Watson mide la independencia de los residuos, el cual toma el valor de 2 cuando los residuos son completamente independientes (entre 1.5 y 2.5 se considera que existe independencia), DW<2 indica autocorrelación positiva y DW>2 autocorrelación negativa" (Universidad de Vigo, 2017, p. 17).

$$DW = \frac{\sum_{i=2}^{n}(e_i - e_{i-1})^2}{\sum_{i=1}^{n} {e_i}^2}, 0 \leq DW \leq 4 \quad (Ec. 12)$$

Donde: e_i se refiere a los residuos ($e_i = Y_i - \hat{Y}_i$).

2.2. Índice oceánico El Niño (ONI) y productividad de caña de azúcar

El índice oceánico El Niño, es un estándar para identificar eventos cálidos (El Niño) y eventos fríos (La Niña) en el océano Pacífico, que consiste en un calentamiento anómalo en gran escala de las aguas superficiales del océano Pacífico central; produce variaciones importantes en las temperaturas y en los patrones pluviales a nivel global con efectos positivos o negativos en la agricultura en general. En la latitud 14°, correspondiente a la zona cañera guatemalteca, El Niño genera mayor incidencia de sequías; mientras que, La Niña mayor incidencia de lluvias, ambos generalmente en condiciones extremas. Se calcula por la media móvil de tres meses de las anomalías de la temperatura

superficial del mar para la región El Niño 3.4, la franja entre 5 °N-5 °S y 120 °E-170 °W; se considera Niño, mayor a 0.5 °C y Niña menor a -0.5 °C (FAO, 2015).

En la producción de caña se ha encontrado alta relación entre el ONI y la productividad. Para el ingenio azucarero en estudio, en los últimos años, las producciones se redujeron en años con efecto de La Niña.

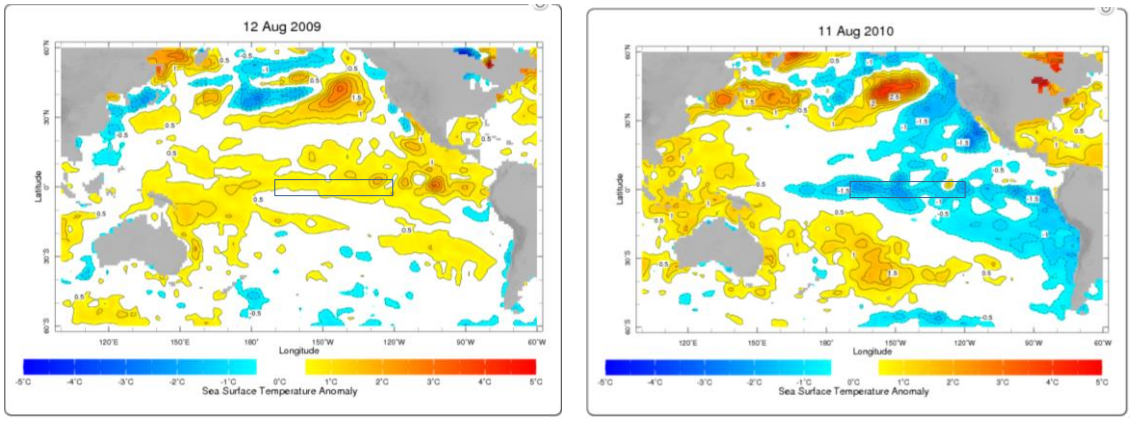
Tabla III. **Rendimiento y productividad del ingenio azucarero y tipo de ENSO durante el período junio-octubre**

ONI	Temperatura°	Año	Toneladas por hectárea	Kilogramos de azúcar por tonelada de caña	Toneladas de azúcar por hectárea
Cálido	0.6	2009/2010	104.66	100.19	10.44
Frío	-1.2	2010/2011	89.11	102.76	9.1
Frío	-0.6	2011/2012	103.85	100.07	10.35
Cálido	0.6	2012/2013	109.42	96.48	10.53
Neutro	-0.4	2013/2014	118.31	95.67	11.25
Neutro	0.0	2014/2015	117.36	103.64	12.14
Cálido	2.0	2015/2016	124.94	97.08	12.13
Frío	-0.6	2016/2017*	112.18	103	11.54

Fuente: elaboración propia.

En la zafra 2010/2011, la producción estuvo 15.55 TH abajo respecto al año anterior, durante el período junio-octubre del año 2010, predominó el fenómeno La Niña; en la zafra 2016/2017 la producción estuvo 12.76 TH abajo respecto a la zafra anterior y corresponde también al efecto del fenómeno de La Niña. Por otro lado, la producción más alta se dio en la zafra 2015/2016 (año Niño, 124.94 TH).

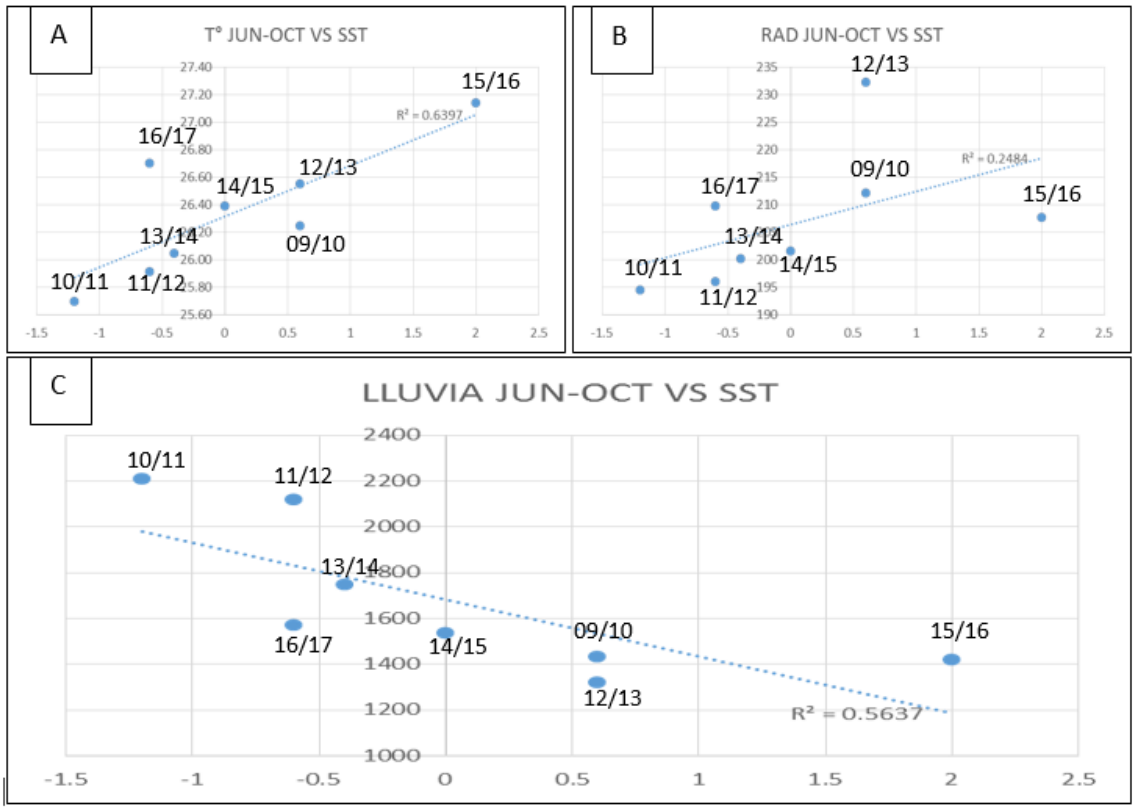
Figura 2. **Oscilación de la temperatura de la superficie del mar a nivel del Ecuador y su efecto en índice oceánico El Niño**



Fuente: FAO. *Entendiendo el impacto de sequía provocada por El Niño en el área agrícola mundial: una evaluación utilizando el Índice de Estrés Agrícola de la FAO (ASI).* Consultado el 15 de marzo de 2020. Recuperado de http://www.fao.org/publications).

En la figura 2, se observa (izquierda) que cuando la oscilación es positiva, mayor a 0.5 °C, se considera un evento cálido (Niño), cuando la oscilación es negativa (derecha), se considera un evento frío (Niña). Cuando un año es frío, en la latitud 14°, correspondiente a la zona cañera guatemalteca, se podría esperar mayor incidencia de lluvias, días más nublados, menor radiación solar acumulada y temperaturas más bajas. Por el contrario, cuando es año cálido, generalmente se esperaría menor incidencia de lluvias, días más soleados, menos nubosidad, mayor radiación solar acumulada y mayores temperaturas (FAO, 2015).

Figura 3. **Relación entre el índice oceánico El Niño y variables climáticas de cada temporada de producción de caña**



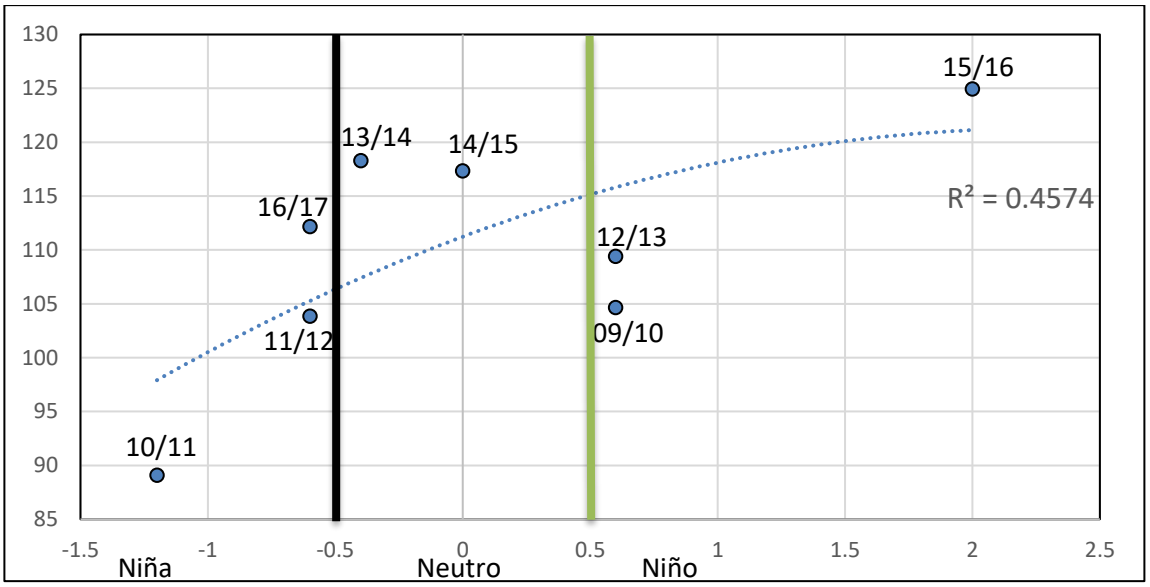
Fuente: elaboración propia.

Cuando el evento cálido fue más intenso, la temperatura media fue mayor; cuando el evento frio fue más intenso, la temperatura media fue menor. Cuando fue año cálido, la radiación tendió a ser mayor, mientras que cuando fue evento frío, la radiación fue menor. En el evento cálido más intenso, la precipitación fue de 1400 mm; mientras que en el año frío más intenso fue de 2200 mm.

Al comparar el efecto del ENSO con la temperatura media de la zona cañera guatemalteca, se observa que, a mayor intensidad del efecto Niño, mayor temperatura media, mientras que mayor intensidad del efecto Niña, menor

temperatura media. Misma tendencia presentó la comparación del ENSO y la radiación solar. Comportamiento contrario se observa en la precipitación pluvial, cuando se tiende hacia año Niña, generalmente la precipitación aumenta, mientras que cuando el efecto de El Niño es mayor, la precipitación pluvial se reduce.

Figura 4. **Relación entre el índice oceánico y la productividad del ingenio azucarero**



Fuente: elaboración propia.

La figura 4 muestra que la oscilación del índice oceánico podría tener efecto en la productividad de caña de azúcar, variando la productividad en función de la intensidad del evento frío o cálido. A medida que el índice fue negativo (mayor influencia del evento frío), la productividad fue menor; a medida que el índice incrementó (mayor influencia del evento cálido), la productividad fue mayor; coincidiendo el evento cálido más intenso en la temporada 2015/2016, con la productividad más alta de los últimos años.

2.3. Índice de vegetación de diferencia normalizada (NDVI) y la productividad de caña de azúcar

Las etapas fenológicas de la caña de azúcar se refieren a la tasa de desarrollo del cultivo (Bonnett, 1998). Para caña de azúcar se han caracterizado 5 etapas fenológicas: a) Iniciación, b) macollamiento, c) elongación 1, d) elongación 2 y e) maduración.

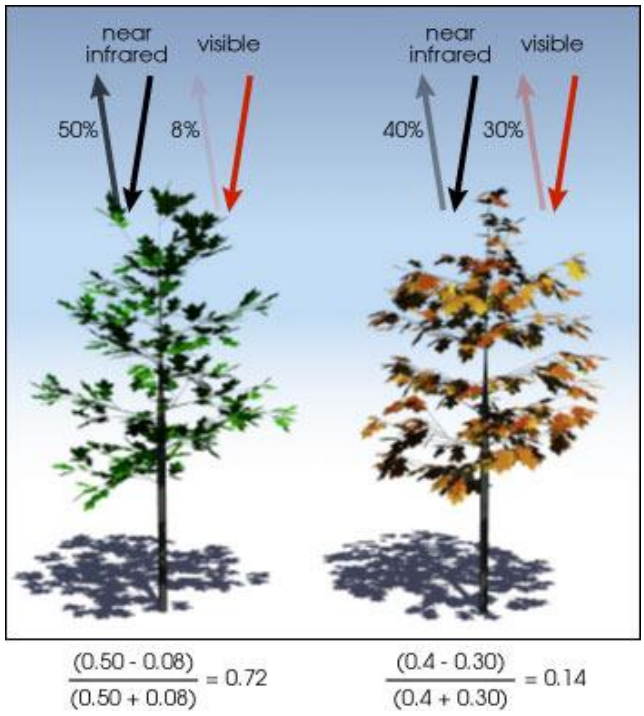
La iniciación, es el período de brotación de la caña de azúcar. El macollamiento es el período donde la caña empieza a desarrollar la macolla, la planta genera la cantidad de tallos que posteriormente entraran a una etapa de competencia para llegar a ser tallos molederos. Elongación 1, es la etapa de máximo crecimiento de la caña de azúcar, los tallos que lograron sobrevivir, empiezan a desarrollarse en grosor, densidad y altura, para convertirse en tallos molederos.

Elongación 2, es una etapa de menor tasa de desarrollo, los tallos formados, siguen con un proceso de crecimiento y desarrollo, pero más lento que la elongación 1. La maduración es la etapa de máxima acumulación de azúcar, los tallos desarrollados y que sobrepasaron el proceso de competencia, entran en la etapa de acumulación de azúcar. Cada una de estas etapas está relacionadas a valores de NDVI.

Un índice de vegetación es un parámetro que unifica en un único valor los datos de múltiples bandas de reflectancia de una imagen y se correlaciona con parámetros de vegetación como biomasa, productividad, índice de área foliar, entre otros (Virginia y Wall, 2001).

Los índices de vegetación se basan en las características espectrales únicas de cada cultivo en las longitudes de onda visible a infrarrojo. El índice de vegetación de diferencia normalizada permite identificar la presencia de vegetación verde en la superficie y caracterizar su distribución espacial y su estado a lo largo del tiempo, a través de la cuantificación del reflejo de la luz en las hojas de las plantas (Virginia y Wall, 2001).

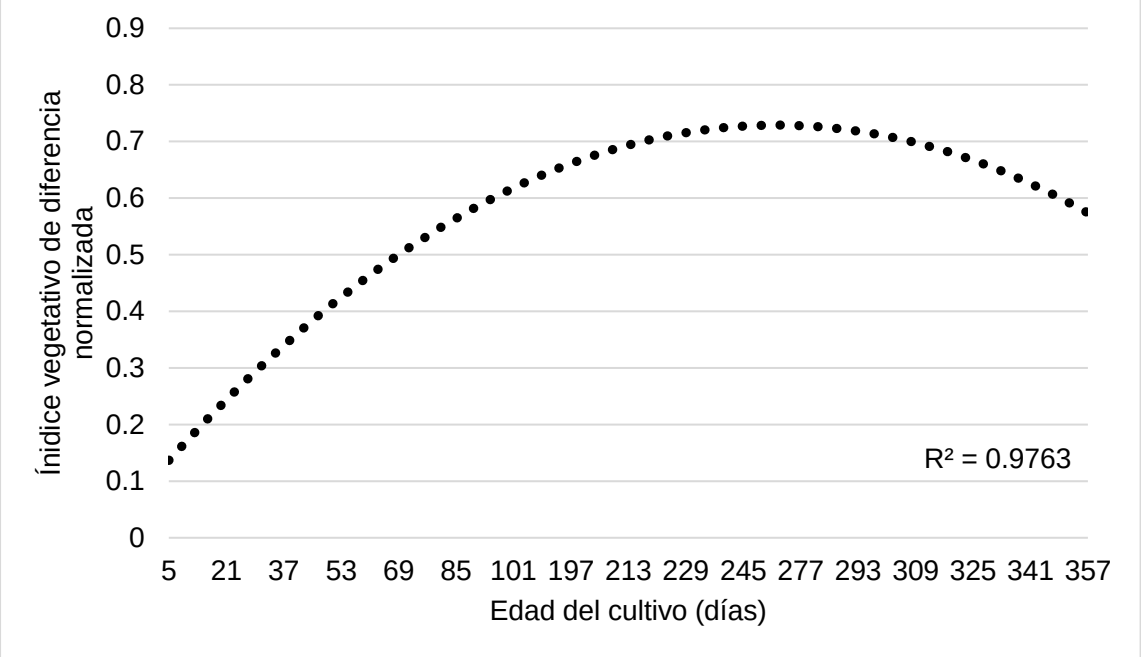
Figura 5. **Diagrama de cálculo de índice NDVI**



Fuente: Virginia y Wall. *Principles of Ecosystem function*. Consultado el 15 de marzo de 2020. Recuperado de https://doi.org/10.1016/j.forpol.2011.08.008.

La figura 5 muestra que a mayor reflectancia de infrarrojo cercano el índice vegetativo de diferencia normalizada aumenta. Alguna variación en la cantidad de reflectancia y absorción de cada una de las distintas ondas se debe a algún cambio metabólico de la planta (Virginia y Wall, 2001).

Figura 6. **Comportamiento del índice vegetativo de diferencia normalizada a lo largo de la edad del cultivo de caña de azúcar**



Fuente: elaboración propia.

El índice vegetativo de diferencia normalizada incrementa a medida que hay mayor desarrollo vegetativo de la caña de azúcar hasta llegar a un máximo en la etapa de elongación 2, se reduce generalmente cuando el cultivo inicia la etapa de maduración, por estrés de tipo hídrico, plagas o deficiencia nutricional (Bégué *et al*., 2010). Se muestra que existe alto grado de ajuste (coeficiente de determinación=0.9763) entre la edad del cultivo y el índice vegetativo de

diferencia normalizada, comportamiento esperado en plantaciones de caña de azúcar con un óptimo desarrollo, a medida que existan factores que reduzcan la capacidad metabólica del cultivo, el desarrollo se frenará y el índice vegetativo de diferencia normalizada será menor.

3. PRESENTACIÓN DE RESULTADOS

3.1. Objetivo 1: Establecer las categorías potenciales de NDVI para la predicción de productividad de caña de azúcar

De acuerdo con los objetivos propuestos por medio de gráficas y tablas se describen todos los indicadores estadísticos utilizados para resolver cada uno de los objetivos planteados.

3.1.1. Recolección de datos de productividad y NDVI

Los datos de productividad de caña de azúcar se obtuvieron por medio del equipo de cosecha mecanizada con el sistema en línea de pesaje de caña, que consistió en una celda de carga colocada en el elevador de la máquina que registró la productividad de caña en toneladas de caña por hectárea, generando un registro de productividad a cada segundo, durante su operación. La caña se depositó en jaulas, según el proceso tradicional de cosecha mecanizada.

Figura 7. **Proceso de captura de datos de productividad**

Fuente: Universidad Autónoma Chapingo. Consultado el 15 de marzo de 2020. Recuperado de https://www.gob.mx/cms/uploads/attachment/file/114363/1.-_Boletin_Julio_2015.pdf.

El sensor se ubicó al final del elevador de caña, se registró la productividad de caña geoespacialmente a cada segundo de operación, se obtuvo un registro de productividad en toneladas de caña por hectárea (TCH) y su coordenada geográfica y se guardaron en la computadora de la cosechadora. A su vez, la cosechadora se encuentra conectada a la red RTK (*Real-time kinematic*). Los receptores RTK son dispositivos que además de captar la señal del satélite GPS, se comunican con otro emisor situado en otro punto fijo (estación base) y realizan cálculos complejos para obtener precisiones cercanas a 1 cm (Schmidt *et al*., 2000). Posteriormente se descargaron los datos y se trasladaron a una computadora para su análisis.

Se utilizó la base de datos de NDVI del satélite *Sentinel*, que tiene una resolución de 10 por 10 metros (100 m^2), para el período del estudio que fue la temporada de producción de 2018-2019. La unidad de análisis fue de 1 lote, cuya área varió entre 15 a 30 hectáreas, posteriormente se enlazaron geoespacialmente las series de datos de NDVI y productividad.

3.1.2. Análisis exploratorio de la base de datos

Para determinar el comportamiento de la serie original, se realizó un análisis exploratorio.

Tabla IV. **Resumen de los 5 números para NDVI y productividad**

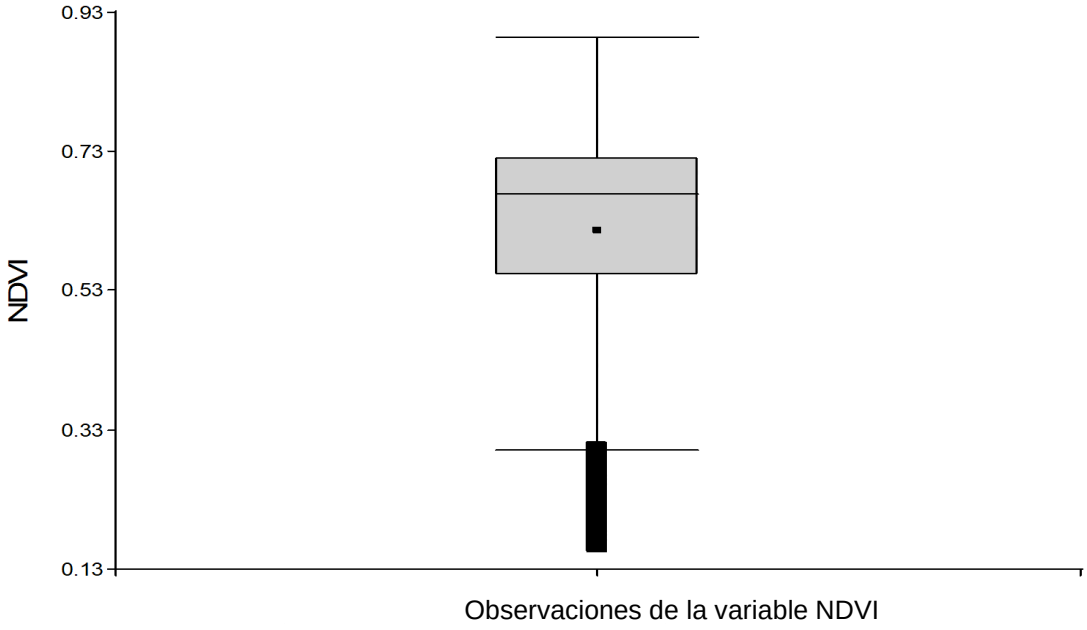
Resumen	NDVI	TCH
N	6345	161
Media	0.58	134.34
D.E.	0.18	22.02
Mín	0.01	38.71
Máx	0.92	188.46
Mediana	0.64	136.82
Q1	0.43	122.05
Q3	0.73	150.34

Fuente: elaboración propia.

Para la serie NDVI se analizaron 6345 datos correspondientes a aproximadamente 4 imágenes por mes, durante el ciclo productivo. El promedio de los datos fue de 0.58, el 50 por ciento de los datos estuvo entre 0.43 y 0.73, el valor mínimo fue de 0.01 y el máximo de 0.92, la mediana fue de 0.64. Para la variable TCH, el número de observaciones fue de 161, correspondiente a los lotes analizados, la media fue de 134.34 TCH, con un valor mínimo de 22.02 y

un máximo de 188.46, el 50 por ciento de los datos estuvieron entre 122.05 y 150.34 TCH.

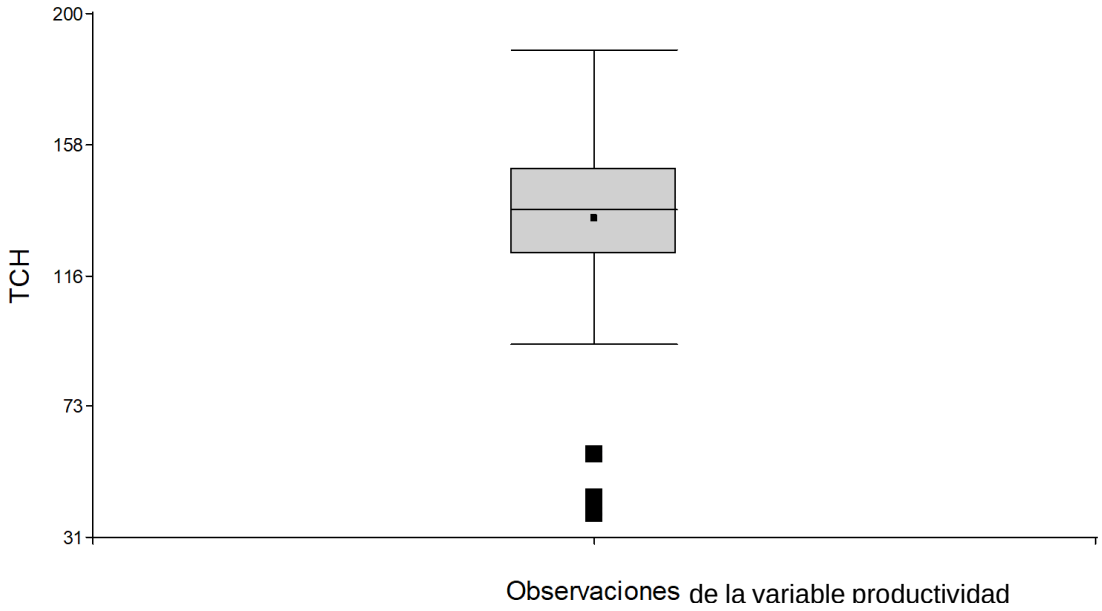
Figura 8. ***Boxplot* para NDVI de la temporada 2018-2019**



Fuente: elaboración propia.

La variable NDVI presentó valores extremos por debajo del primer cuartil, estos valores se contrastaron con la variable productividad, para entender el comportamiento de ambas variables.

Figura 9. ***Boxplot* para productividad de la temporada 2018-2019**



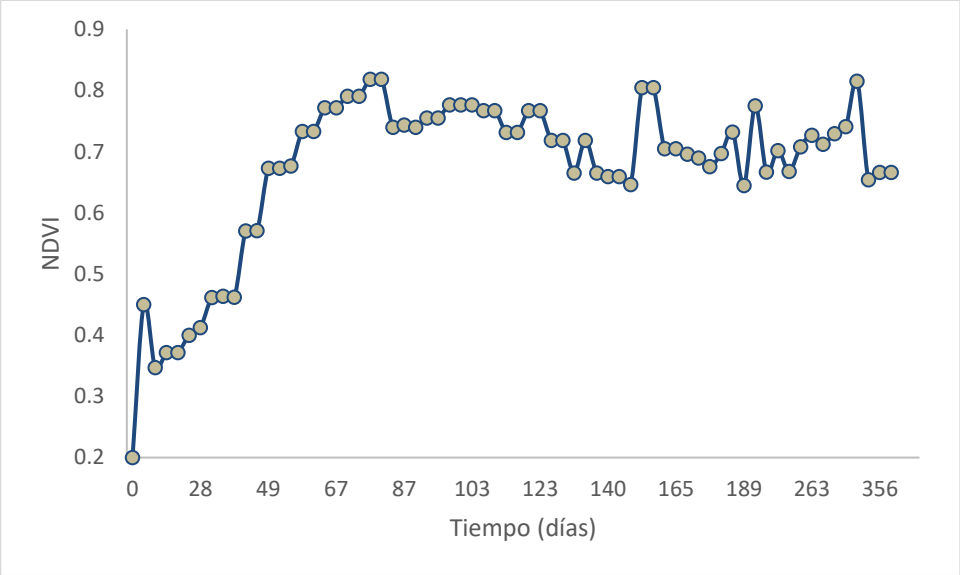
Fuente: elaboración propia.

Se observa que la variable productividad, presentó valores por debajo del primer cuartil diferentes al resto, para este estudio estos valores extremos sirvieron para entender el comportamiento del NDVI, contrastado con valores de productividad.

3.1.3. Identificación de categorías como variables regresoras para el análisis de regresión

Para este estudio, el ciclo de producción de caña de azúcar fue considerado de 12 meses. En Guatemala, el ciclo de producción de caña de azúcar inicia en noviembre, durante la época seca y finaliza en abril del siguiente año.

Figura 10. **Ejemplo de serie de NDVI, desde 0 hasta 365 días, para las condiciones de región cañera de Guatemala**



Fuente: elaboración propia.

Para las condiciones de la región cañera de Guatemala, donde el cultivo comienza a crecer en la época de verano (seca), a partir de noviembre y su ciclo tiene una duración de 365 días, el NDVI muestra un comportamiento ondulante con altas y bajas, dependiendo de la variabilidad climática, que se traduce en incremento o disminución de la tasa de respiración de la planta, lo que está relacionado a la producción de biomasa.

Acorde a Lobato, Favilla y Mora (2017), se categorizaron los datos con 21 características, con 20 características espectrales independientes en función del NDVI y una característica dependiente para predecir (TCH), los valores de NDVI, correspondieron al promedio de los pixeles que se encontraron dentro del lote de producción.

Tabla V. **Categorías generadas para la serie de tiempo de NDVI**

Abreviatura de categoría	Descripción de categoría
Mes1	El promedio de las imágenes obtenidas desde el corte hasta los primeros 30 días de desarrollo del cultivo.
Mes2	El promedio de las imágenes obtenidas desde el día 31 hasta los primeros 60 días de desarrollo del cultivo.
Mes3	El promedio de las imágenes obtenidas desde el día 61 hasta los primeros 90 días de desarrollo del cultivo.
Mes4	El promedio de las imágenes obtenidas desde el día 91 hasta los primeros 120 días de desarrollo del cultivo.
Mes5	El promedio de las imágenes obtenidas desde el día 121 hasta los primeros 150 días de desarrollo del cultivo.
Mes6	El promedio de las imágenes obtenidas desde el día 151 hasta los primeros 180 días de desarrollo del cultivo.
Mes7	El promedio de las imágenes obtenidas desde el día 181 hasta los primeros 210 días de desarrollo del cultivo.
Mes8	El promedio de las imágenes obtenidas desde el día 211 hasta los primeros 240 días de desarrollo del cultivo.
Mes9	El promedio de las imágenes obtenidas desde el día 241 hasta los primeros 270 días de desarrollo del cultivo.
Mes10	El promedio de las imágenes obtenidas desde el día 271 hasta los 300 días de desarrollo del cultivo.
Mes11	El promedio de las imágenes obtenidas desde el día 301 hasta los 330 días de desarrollo del cultivo.
Mes12	El promedio del período del día 331 hasta los 365 días de desarrollo del cultivo.
ini	el promedio de las imágenes obtenidas desde el corte hasta los primeros 45 días.
mac	El promedio de las imágenes obtenidas desde el día 45 hasta los primeros 135 días de desarrollo del cultivo.
elo1	El promedio de las imágenes obtenidas desde el día 136 hasta los primeros 215 días de desarrollo del cultivo.
elo2	El promedio de las imágenes obtenidas desde el día 216 hasta los primeros 300 días de desarrollo del cultivo.

Continuación de tabla V.

mad	El promedio de las imágenes obtenidas desde el día 301 hasta los primeros 365 días de desarrollo del cultivo.
acuinma	La suma del promedio del mes1 y mes2.
acuinmael1	La suma del promedio del Mes1, Mes2 y Mes3.
acuinmael12	La suma del promedio del Mes1, Mes2, Mes3 y Mes4.

Fuente: elaboración propia.

Se realizó un análisis del coeficiente de correlación lineal de Pearson, para determinar las categorías que podrían tener correlación significativa con la productividad e identificar las adecuadas para incluir en la generación de modelos para las categorías de meses.

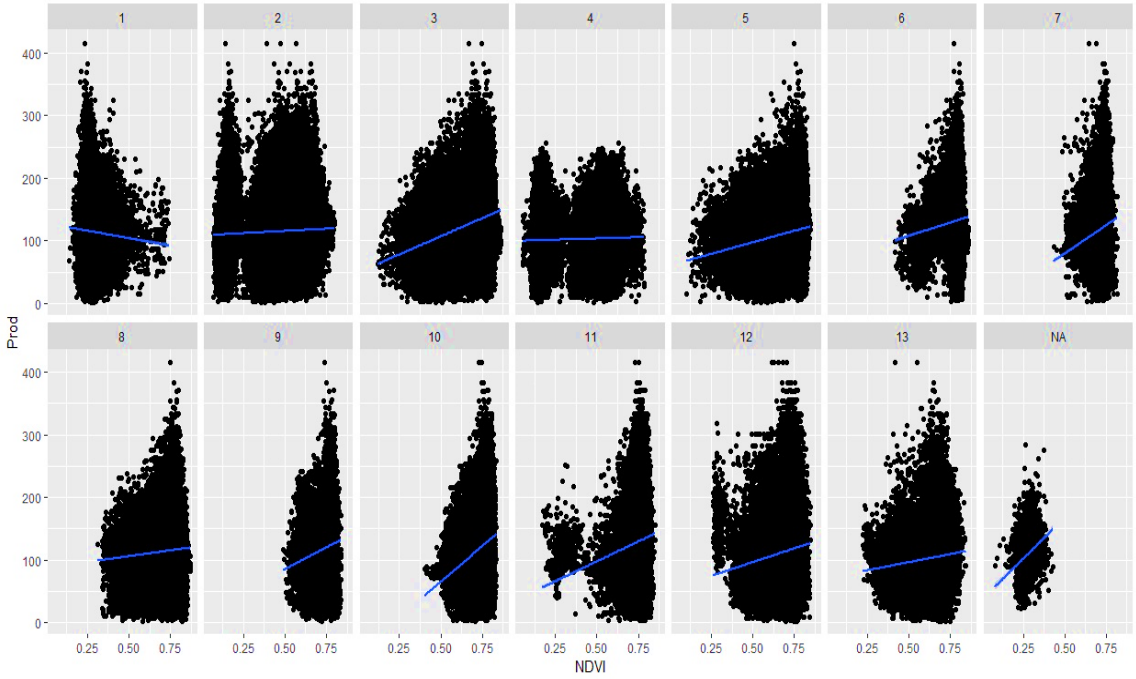
Tabla VI. **Correlación de Pearson en categorías de NDVI y productividad**

Correlación Pearson		Valor de *p*
Categoría de NDVI	Productividad	
Mes1	-0.05	0.85
Mes2	-0.02	0.95
Mes3	0.56	0.10
Mes4	0.51	0.12
Mes5	0.60	0.10
Mes6	0.30	0.21
Mes7	-0.18	0.46
Mes8	-0.32	0.18
Mes9	0.66	0.01
Mes10	0.73	0.01
Mes11	0.45	0.06
Mes12	-0.15	0.59

Fuente: elaboración propia.

En general, las categorías de meses mostraron un coeficiente de correlación lineal de Pearson de bajo grado de asociación (desde r= 0.01 hasta r=0.59), entre nula asociación y baja correlación positiva. Solamente se observa el Mes1 y Mesr2, con un coeficiente alto de r=0.85 y r=0.95.

Figura 11. **Diagrama de dispersión entre las categorías de meses y la productividad**



Fuente: elaboración propia.

Aunque el coeficiente de correlación lineal de Pearson mostró alguna correlación positiva entre el NDVI mensual y la productividad, en las categorías Mes1 y Mes2, las bajas correlaciones se reflejan en el diagrama de dispersión. El coeficiente de correlación lineal de Pearson pudo ser afectado por la gran cantidad de observaciones dentro de cada mes.

Se realizó el mismo procedimiento para las categorías de etapas fenológicas, se calculó el coeficiente de correlación de Pearson y se obtuvieron valores de alta correlación, que se describen en la tabla VII.

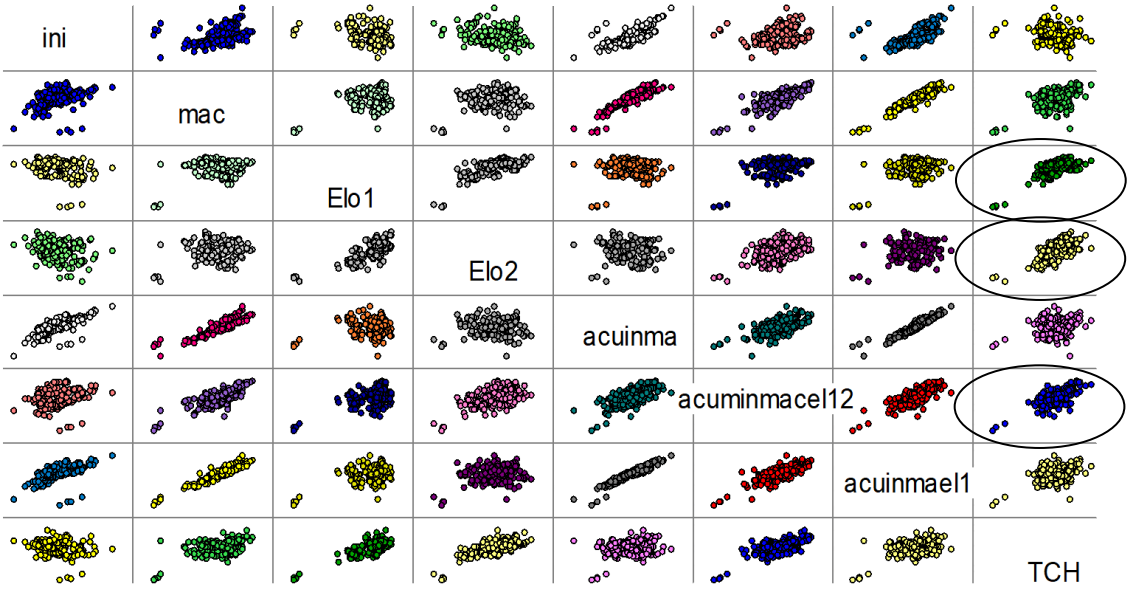
Tabla VII. **Correlación lineal de Pearson entre categorías de NDVI por etapa fenológica y productividad**

Correlación de Pearson		Valor *p*
Categorías de NDVI-TCH	Pearson	
Ini	-0.25	0.015
Mac	0.58	0.00
elo1	0.79	0.00
elo2	0.72	0.00
Mad	0.33	0.00
Acuinma	0.45	0.00
acuiminmacel12	0.68	0.00
acuinmael1	0.70	0.00

Fuente: elaboración propia.

El coeficiente de Pearson mostró correlación alta positiva entre las categorías elongación 1, elongación 2, acuinmacel12 y acuinmael1.

Figura 12. **Matriz de correlación para las categorías de estaciones fenológicas y la productividad**



Fuente: elaboración propia.

Para corroborar los valores del coeficiente de correlación lineal de Pearson, se realizó la matriz de correlación gráfica, se observó que existe correlación entre la productividad y elongación 1, elongación 2, acuminmacel12 (abreviación para la sumatoria del promedio de iniciación, macollamiento y elongación 1 y elongación 2).

El mayor grado de asociación se dio en las categorías de etapas fenológicas (5 etapas fenológicas), comparado con las categorías de meses (12 meses). Este comportamiento puede explicarse con el teorema del límite central, que indica que las medias de las medias de muestras grandes y aleatorias son aproximadamente normales. Para este caso, las etapas fenológicas resumieron los 365 días del cultivo sumarizando en 5 grupos la serie de tiempo de NDVI e

incrementando el grado de asociación de las variables regresoras potenciales y la productividad.

Se utilizaron las categorías de etapas fenológicas para generar todos los modelos posibles, maximizando el coeficiente de determinación R^2 y minimizando el índice de información de Akaike (AIC).

3.2. Objetivo 2: Seleccionar las variables regresoras (categorías) que contribuyan a obtener un mayor grado de ajuste del modelo

Después de obtener las variables potenciales que fueron principalmente aquellas relacionadas a la agrupación por etapa fenológica del cultivo, se procedió a realizar la selección de las variables regresoras que pueden contribuir a obtener un modelo de mayor grado de ajuste.

3.2.1. Modelos de regresión maximizando el coeficiente de determinación R^2 y minimizando el índice de información de Akaike (AIC)

Se realizó la modelación de regresión, utilizando las 7 variables provenientes de la etapa fenológica y se calcularon todos los modelos, por medio del método de maximización del R^2 y minimización del AIC.

Tabla VIII. **Primera generación de modelos de regresión**

R2	Variable 1	Variable 2	Variable 3	Variable 4	Variable 5
0.68	elo1	mad	acutot	acuinelo1mad	
0.68	elo1	mad	acuminmacel12	acuinelo1mad	
0.68	Ini	elo1	Acutot	acuinelo1mad	

Continuacion tabla VIII.

0.68	Ini	elo1	acuminmacel12	acuinelo1mad	
0.68	Ini	elo1	Mad	acuminmacel12	
0.68	Ini	elo1	Mad	acutot	
0.68	Ini	mad	Acutot	acuinelo1mad	
0.68	Ini	mad	acuminmacel12	acuinelo1mad	
0.68	Ini	elo2	acuinmael1	acuinelo1mad	
0.68	mac	elo1	Elo2	acuinmael1	acuinelo1mad

*485 modelos estimados no mostrados, por R^2 menor a los primeros 10

Fuente: elaboración propia.

Se calcularon 495 modelos con las variables regresoras de mayor correlación, los mejores modelos generados, fueron los de regresión múltiple, de 4 variables regresoras.

Se procedió a generar el modelo de regresión con el mejor modelo que maximizó el R^2 y minimizó el AIC.

Tabla IX. **Análisis de regresión lineal**

Variable	N	R^2	R^2 Aj	ECMP	AIC	BIC
TCH	170	0.69	0.68	179.02	1355.54	1374.36

Fuente: elaboración propia.

El modelo generado presentó un R^2 de 0.69 y un AIC de 1355.54, este modelo fue el que obtuvo el máximo coeficiente de determinación y el valor mínimo del índice de determinación de Akaike.

Tabla X. **Coeficientes de regresión y estadísticos asociados**

Coef	Est	E.E.	LI (95 %)	LS (95 %)	T	valor P
Const	-210.51	27.74	-265.27	-155.74	-7.59	<0.0001
Elo1	272.58	32.88	207.67	337.5	8.29	<0.0001
Mad	115.12	25.78	64.23	166.02	4.47	<0.0001
acutot	110.99	19.99	71.51	150.46	5.55	<0.0001
acuinelo1mad	-158.71	46.05	-249.64	-67.78	-3.45	<0.0001

Fuente: elaboración propia.

Las cuatro variables del modelo de regresión resultaron significativas.

Tabla XI. **Análisis de varianza para el modelo de regresión**

Fuente de variación	Suma de cuadrados	grados de libertad	Cuadrado medio	F	Valor
Modelo	60100.69	4	15025	92.05	<0.0001
elo1	11221.93	1	11221.93	68.75	<0.0001
Mad	3255.67	1	3255.67	19.95	<0.0001
Acutot	5030.38	1	5030.38	30.82	<0.0001
acuinelo1mad	1938.72	1	1938.72	11.88	<0.0001
Error	26933.27	165	163.23		
Total	87033.97	169			

Fuente: elaboración propia.

El análisis de varianza determinó que las cuatro variables regresoras son significativas en el modelo, es decir, ninguna de las cuatro regresoras generó colinealidad en el modelo y aportan en el pronóstico.

De las cuatro variables regresoras, la que más explica la variabilidad de la productividad es la etapa de elongación 1 (elo1, suma de cuadrados 3255.67),

porque la etapa de elongación 1, es el período comprendido entre los 135 a 250 días de desarrollo del cultivo, es la etapa donde el cultivo tiene la tasa de crecimiento más alta y la etapa de mayor cantidad de días de desarrollo del cultivo.

Aunque el primer modelo cumple con un buen ajuste R^2=0.69, se necesitó esperar hasta los 320 días para realizar el pronóstico de la producción.

Tomando en cuenta que la variable de mayor peso fue Elongación 1, que se obtiene a los 135 días de desarrollo del cultivo, se volvieron a generar modelos, eliminando aquellas categorías con más de 300 días de desarrollo del cultivo, para obtener un modelo que antes de los 300 días pudiera proyectar la producción del lote.

Tabla XII. **Segundo grupo de modelos de regresión, con variables de menor edad**

R^2	Variable 1	Variable 2	Variable 3	Variable 4	Variable 5
0.75	mac	Elo1	acuinmael1		
0.75	mac	Elo1	acuinma	acuminmacel12	
0.75	Mac	acuinma	acuinmael1	acuminmacel12	
0.74	Mac	Elo1	acuinma	acuinmael1	acuminmacel12
0.73	Elo1	acuinmael1	acuminmacel12		
0.73	Elo1	acuinma	acuminmacel12		
0.73	acuinma	acuinmael1	acuminmacel12		
0.72	elo1	acuminmacel12			
0.72	mac	Elo1	acuminmacel12		
0.69	mac	Elo1	acuinmael1		
*2 modelos que no se presentan, con R^2 menor a 0.69					

Fuente: elaboración propia.

El segundo grupo de modelos generados muestra una mejora en los niveles de R^2, pasando a obtener valores de 0.75 y categorías de menor edad del cultivo, hasta elongación 2.

3.3. Objetivo 3: Determinar el grado de ajuste del modelo que mejor representa la relación entre la productividad y la o las variables regresoras

Después de obtener las variables regresoras que aportan al modelo, se realizó el análisis de regresión para obtener los estadísticos de cada modelo.

3.3.1. Modelo de mejor ajuste para pronosticar la productividad de caña de azúcar

El modelo que maximizó el R^2 y minimizó el AIC fue generado con las variables regresoras macollamiento, elongación 1, el acumulado (suma) de iniciación, macollamiento y elongación 1.

Tabla XIII. **Análisis de regresión lineal de modelo de mejor ajuste**

Variable	N	R^2	R2 Aj	ECMP	AIC	BIC
TCH	170	0.75	0.75	128.73	1236.59	1252

Fuente: elaboración propia.

Comparado con el mejor modelo del primer grupo de modelos de regresión, el segundo modelo, presentó un menor índice de Akaike, o la menor verosimilitud (1236.59), comparado con el primero que presentó un índice de Akaike de 1374. Además, el segundo modelo cumple con el principio de Parsimonia, que

establece que la solución más simple suele ser la mejor, en este caso, el segundo modelo logró reducir el número de variables regresoras a tres.

Tabla XIV. **Coeficientes de regresión y estadísticos asociados**

Coef	Est	E.E.	LI (95%)	LS (95%)	T	valor P
Const	-135.28	21.17	-177.09	-93.47	-6.39	<0.0001
Mac	271.93	36.22	200.4	343.46	7.51	<0.0001
el1	410.29	25.4	360.12	460.47	16.15	<0.0001
acuinmael1	-117.31	23.92	-164.56	-70.05	-4.9	<0.0001

Fuente: elaboración propia.

Las 3 variables fueron significativas en el modelo (valor P <0.0001), por lo que no existe colinealidad en las variables y son importantes para la predicción de la productividad.

Tabla XV. **Análisis de varianza para el modelo de regresión**

Fuente de variación	Suma de cuadrados	grados de libertad	Cuadrado medio	F	Valor P
Modelo	58357.4	3	19452.47	159.2	<0.0001
mac	6890.84	1	6890.84	56.38	<0.0001
elo1	31884.66	1	31884.66	260.9	<0.0001
acuinmael1	2938.88	1	2938.88	24.04	<0.0001
Error	19189.23	157	122.22		
Total	77546.62	160			

Fuente: elaboración propia.

Comparado con el primer modelo, el segundo modelo mostró que la variable que más explica la variabilidad de la productividad fue la etapa de elongación 1,

pero en el segundo modelo esta variabilidad explicada fue mayor 41.16 %, contra 12.9 % del primer modelo. Se observó también que la variabilidad inherente al error se redujo en el segundo modelo.

El modelo que maximizó el ajuste y presentó el menor índice de verosimilitud para el pronóstico de la productividad de caña de azúcar, a los 7 meses de desarrollo del cultivo fue:

$$Productivida = 271.9(macollamiento) + 410.3(elongación1) - 117.3(acuinmacelo1) - 135.28$$

Donde: productividad TCH es igual a macollamiento (NDVI promedio del día 46 al día 145), más elongación 1 (igual al promedio del NDVI del día 146 al día 215), más acuinmacelo1 (igual a la suma de las medias de NDVI de iniciación, macollamiento y elongación 1).

3.4. Objetivo general: Determinar el modelo de mejor ajuste, utilizando el Índice vegetativo de diferencia normalizada (NDVI), que proyecta la producción del ingenio azucarero

Para resolver el problema planteado, se presenta el objetivo general

3.4.1. Validación de los supuestos del modelo

Se evaluaron los supuestos del análisis de regresión para el mejor modelo

3.4.1.1. Validación de la normalidad de los residuos del modelo por *Kolmogov Smirnov*

Para la validación de la normalidad, se utilizó la prueba de bondad de ajuste Kolmogorov Smirnov, porque la cantidad de datos utilizados fue superior a 100 muestras y dicha prueba no pierde potencia con muestras mayores a 100.

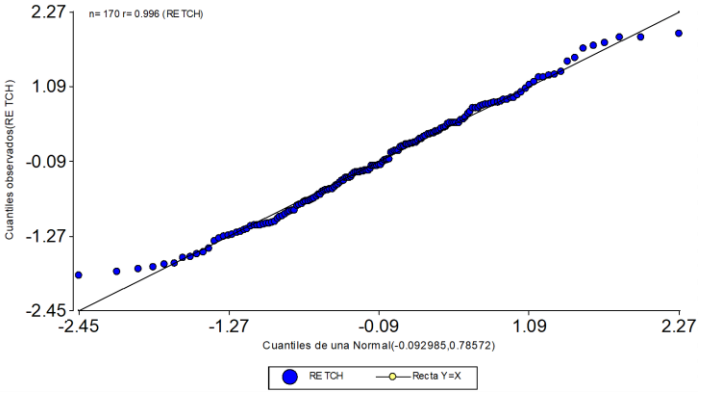
Tabla XVI. **Prueba de bondad de ajuste Kolmogorov Smirnov para normalidad de los residuos del modelo**

Variable	Ajuste	Media	Varianza	Estadístico D	P-valor
RE TCH	Normal (0-1)	-0.09	0.79	0.07	0.4725

Fuente: elaboración propia.

Con un porcentaje de confianza mayor a 95 % la prueba Kolmogorov Smirnov, determinó que la distribución de los residuos sigue una distribución normal.

Figura 13. ***QQplot* para los residuos de productividad**



Fuente: elaboración propia.

El *QQplot* efectivamente muestra que la distribución de los residuos del modelo se ajusta a una distribución normal.

3.4.1.2. Heterocedasticidad por Breusch-Pagan

La prueba de Breusch-Pagan se utilizó para probar la heterocedasticidad en un modelo de regresión.

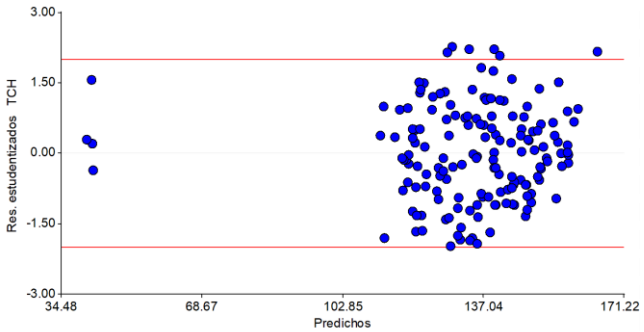
Tabla XVII. **Prueba para heterocedasticidad por Breusch-Pagan**

GL	Estadístico BP	P-valor
3	6.5576	0.08742

Fuente: elaboración propia.

Determina si la varianza de los errores de una regresión depende de los valores de las variables independientes. Para el modelo de mejor ajuste, el estadístico BP determinó que se acepta la hipótesis nula de homocedasticidad de varianzas.

Figura 14. **Correlograma de predichos y residuos estudentizados**



Fuente: elaboración propia.

En el correlagrama se observa que no existe ningún patrón de comportamiento de los predichos respecto a los residuos estudentizados y la mayoría de las observaciones estuvieron dentro de los límites. Por lo tanto, se puede mantener que estas variables aleatorias no están correlacionadas.

3.4.1.3. Independencia por Durbin-Watson

Este supuesto requiere que la probabilidad de que el residuo de n observaciones tenga un valor particular, no debe depender de los valores de los otros residuos, para tener certeza del cumplimiento de este supuesto se realizó la prueba Durbin-Watson:

Tabla XVIII. **Test Durbin-Watson para independencia de residuos**

Autocorrelación	Estadístico D-W	p-valor
0.4094536	1.179047	0

Fuente: elaboración propia.

La ejecución de la prueba de residuos independientes para el modelo determinó que los residuos no están correlacionados, debido a que el DW está próximo a 2 y el valor de p (p-value) es superior al nivel de significancia de 5 % (alfa=0.05) por lo que se concluye que existe independencia de los residuos.

4. DISCUSIÓN DE RESULTADOS

Por medio del estudio de la serie de tiempo de NDVI, con la aplicación de técnicas de análisis de regresión, se logró obtener un modelo funcional con un AIC de 1236.59 y R^2=0.75, que predice la productividad de caña de azúcar a partir de 7 meses de edad del cultivo. Los resultados fueron plasmados en gráficas y tablas para lograr una lectura de fácil interpretación y compresión.

4.1. Análisis Interno

A continuación, se aborda la interpretación de los resultados obtenidos durante el proceso planteado para el cumplimiento de los objetivos. En cuanto al objetivo general, se logró un modelo objetivo y preciso que explica el 75 % de la variabilidad de la productividad de caña de azúcar, a partir del 7 mes de edad del cultivo, lo que se traduce en menor incertidumbre para la planificación de materiales, reajustes en las proyecciones de cuotas de entregas de azúcar, reajustes en disponibilidad de maquinaria para cosecha.

Durante el desarrollo de la investigación se presentaron varios aspectos positivos como el apoyo y respaldo de la empresa para la realización del trabajo investigativo, el acompañamiento y aportes oportunos del gerente de investigación y desarrollo en todo el proceso de desarrollo del modelo de regresión. Otro aspecto positivo es que los resultados del modelo generan una alarma en el administrador del cultivo, porque al obtenerse una predicción por debajo de lo esperado, el equipo agrícola inicia a recopilar información para validar dicho resultado y entender las causas que podrían generar la merma de

producción, generando un proceso de mejora continua directamente en la gestión agrícola.

Entre los aspectos negativos, el modelo logra una predicción adecuada a partir del séptimo mes de edad, dejando solamente 5 meses de margen para realizar correcciones en el cultivo y en toda la cadena de valor, a partir del séptimo mes hacia atrás, el ajuste del modelo se redujo.

4.2. Análisis externo

Respecto a la metodología utilizada en el presente estudio se puede mencionar que es muy valiosa la experiencia y conocimiento que el investigador tenga sobre el desarrollo del cultivo de caña de azúcar, las etapas fenológicas, tasas de crecimiento, porque facilita la selección de variables potenciales regresoras para el modelo. Los aspectos importantes de resaltar sobre la construcción del modelo de regresión y sus diferencias y similitudes con otros estudios son los siguientes:

4.3. Análisis exploratorio de la base de datos.

Para la serie NDVI, el *boxplot* mostró que existen algunos pocos valores bajos, tan bajos que colocan el promedio por debajo de la mediana, indicando un sesgo negativo. Schneider, Hadad y Kemerer (2013), segmentaron la serie de NDVI en 2 categorías caña caída (caña tirada y volcada dentro del lote) y caña en pie (caña erecta en su posición natural), con el objetivo de encontrar relaciones con el NDVI, los autores reportaron un corrimiento hacia valores inferiores en la serie de datos NDVI en caña de azúcar en pie, y un corrimiento hacia valores superiores en la caña caída.

En la zona cañera guatemalteca, se seleccionan variedades resistentes al volcamiento por vientos, porque es un factor de reducción de productividad (CENGICAÑA, 2016). Para la serie de NDVI bajo análisis se observó un comportamiento parecido a la que reportan los autores mencionados, para caña en pie, indicando que son datos promisorios, porque se reduce la probabilidad de errores al utilizar lotes que durante su período de crecimiento fueran afectados por eventos de vientos fuertes u otros factores de volcamiento, es decir, fueron lotes con un desarrollo normal en la mayoría de su ciclo productivo.

Hernandez, Escribano y Tarquis (2014), evaluaron el comportamiento del NDVI en diversos pastos (familia *Poaceae*), para determinar un rango espacial homogéneo, ellos determinaron que, para los pastos evaluados, el valor modal fue 0.56, mientras que el valor promedio fue 0.41, concluyendo que los pastos evaluados en Salamanca mostraron una clara asimetría entre las colas. El mismo comportamiento reflejaron los datos usados para este estudio, la media estuvo por debajo de la moda y con un sesgo negativo, indicando que fueron datos correspondientes a un desarrollo normal del cultivo de caña de azúcar.

El *boxplot* de productividad mostró un comportamiento menos asimétrico, aunque existieron valores por muy por debajo, la media aritmética no se separó de la mediana, mostrando una distribución más uniforme. Los datos inferiores de productividad pueden estar asociados a lotes alta incidencia de espacios vacíos, sin acceso a riego o fertilización.

4.4. Categorías potenciales de NDVI para la predicción de productividad de caña de azúcar

La serie de tiempo de NDVI, para la zona cañera guatemalteca, tiene una duración entre 11.5-12.5 meses, para este estudio se definió un ciclo de cultivo

de 12 meses, durante este período, la serie de tiempo de NDVI, va registrando altos y bajos, correspondiente a variaciones en el metabolismo de la planta, efectos climáticos y efectos de manejo.

La serie NDVI se compuso de una secuencia de fotografías (40-52 fotografías) por lote que plasmó la condición del cultivo de ese momento. La estimación de productividad a partir del NDVI ha sido conducida a varias escalas, a nivel de país, de región, de temporada, a nivel de ingenio (Lobato, Vieira y Camargo, 2011). Para este estudio se estimó la variabilidad a nivel de lote de producción del ingenio, que abarca un área entre 15 a 35 hectáreas y es la unidad mínima para realizar los presupuestos en el ingenio azucarero.

Pinheiro *et al*., (2018) estudiaron el NDVI como una serie de tiempo para predecir la productividad de caña de azúcar a escala de unidad de área. Para sus estudios, los autores, segmentaron la serie de datos en períodos de tiempo, agrupando a nivel mensual, anual, máximos y mínimos, acumulados y etapas fenológicas, entre otros.

Fernandes *et al,* evaluaron 51 categorías de NDVI, agrupadas en 3 fases. La fase 1, consistió en categorías en función de mínimos, máximos, promedios aritméticos y sumatorias de dichas categorías. La fase 2, consistió en mínimos, máximos y promedios aritméticos de NDVI y mínimos, máximos y promedios aritméticos de temperatura. La fase 3, consistió en la categorización en función del ciclo de cultivo, segmentando por mes y etapas de cultivo.

Los autores evaluaron las categorías y determinaron que los mejores coeficientes de determinación los obtuvieron 2 modelos de regresión simple, el primero utilizando la categoría ndvifinal (valor de los últimos 10 días de NDVI

antes de la cosecha) con un R^2 de 0.561. El segundo modelo con la categoría ndvi2_m (promedio aritmético de valores de la fase 3).

En otro estudio, Fernades, Favilla y Mora (2017), segmentaron la serie NDVI en 20 categorías regresoras que fueron excluyendo por medio de un análisis de eliminación secuencial hacia atrás, los autores determinaron que el modelo de mejor ajuste utilizó las regresoras *m* (valor de NDVI en el medio del ciclo de cultivo), *s* (inicio del ciclo de cultivo) y *e* (final del ciclo de cultivo); con un R^2 de 0.60.

Lofton *et al*., (2012), relacionaron el NDVI acumulado en el ciclo de cultivo con la concentración de microelementos en la hoja para la predicción de productividad, obteniendo modelos de regresión múltiple con coeficientes de determinación entre 0.76 a 0.94.

En este estudio se evaluaron 21 categorías de NDVI, relacionadas a la temporalidad de la serie, segmentada por mes y por etapa fenológica, que, basado en los estudios descritos, han demostrado más grado de asociación con la variable de productividad de caña de azúcar.

Por medio del coeficiente de correlación lineal de Pearson, se evidenció que las variables asociadas a meses no lograron un grado de asocio significativo, aunque algunas categorías obtuvieron un p valor significativo, la prueba pudo ser afectada por la cantidad de la muestra, los grados de libertad pudieron maximizar el valor crítico, porque aunque las categorías Mes9 y Mes10 fueron significativas, por lo tanto se puede concluir que ninguna de las poblaciones evaluadas es distinta de cero, además el gráfico de dispersión mostró que el grado de ajuste es muy bajo.

La matriz de dispersión entre la productividad y las categorías de NDVI por etapas fenológicas, mostró mejor grado de asociación entre las categorías, el cálculo del coeficiente de correlación lineal de Pearson confirmó que las categorías de etapas fenológicas obtuvieron un mayor grado de ajuste, comparado con las categorías de meses, porque los coeficientes de correlación fueron mayores y todas las categorías significativas. Coincidiendo con lo reportado por Lofton *et al.*, (2012), quienes determinaron que las mejores relaciones de los modelos generados estuvieron asociados a las categorías que agruparon períodos grandes del ciclo de cultivo.

Por lo tanto, se determinó en base al coeficiente de correlación lineal de Pearson y el valor P, que las categorías de NDVI por etapa fenológica son las categorías potenciales como regresoras que explican la variabilidad de la productividad de caña de azúcar porque presentaron grado de asociación alta positiva y fueron significativos.

4.5. Selección de las variables regresoras (categorías) que contribuyeron a obtener un mayor grado de ajuste del modelo

Se realizó el análisis de regresión tomando en cuenta las 7 categorías potenciales, relacionadas a las etapas fenológicas para la predicción de productividad de caña de azúcar.

Se generaron 495 modelos, bajo el método de maximización del coeficiente de determinación (R^2) y minimización del índice de información de Akaike. (AIC). En la tabla VIII, se observó que existieron 8 modelos con el mismo R^2, pero solamente el primer modelo minimizó el AIC. Akaike (1985) concluyó que “el estadístico AIC es una medida de desajuste de la distribución predictiva al

especificar el modelo real, y que la minimización de este índice supone minimizar el desajuste y obtener el mejor modelo predictivo" (p. 723).

Por lo tanto, como criterio principal se tomó AIC y las variables regresoras que minimizaron el AIC fueron: elongación 1 (elo1), maduración (mad), el acumulado de iniciación, macollamiento, elongación 1, elongación 2 y maduración (acutot), el acumulado de iniciación, elongación 1 y maduración (acuinelo1mad). Se generó el modelo de regresión múltiple con las mejores regresoras y se obtuvo el modelo:

$$Productividad = 272.58(Elo1) + 115.12(Mad) + 110.99(acutot) - 158.71(acuinelo1mad) - 210.51$$

Donde elo1 es elongación 1, mad es maduración, acutot es el acumulado de iniciación, macollamiento, elongación 1, elongación 2 y maduración, acuinelo1mad es el acumulado de iniciación, elongación 1 y maduración.

El modelo obtuvo un R^2 de 0.69, que indica un alto grado de ajuste entre las regresoras, explicando el 69 % de la variabilidad de la productividad y minimizó el AIC a 1374.36. En el modelo generado, la constante y la variable acuinelo1mad están restando en el modelo, mientras que la variable Elo1, Mad y acutot suman a la variabilidad total de la productividad. Todas las variables fueron significativas para el modelo, por lo tanto, se consideran importantes y no generan colinealidad.

El análisis de varianza del modelo indicó que todas las regresoras fueron significativas. Sin embargo, de las 7 fuentes de variación, el error representa el 30.1 % de la variabilidad de la productividad. De las 4 variables regresoras, la

de mayor impacto en el modelo fue elongación, representando el 13 % de la variabilidad de la productividad.

Aunque el modelo generado cumple con un buen grado de ajuste y valores significativos de las 4 variables regresoras, no es funcional para el pronóstico de la productividad de caña de azúcar porque se necesitaría que el cultivo alcance la etapa de maduración, es decir, 320 días de desarrollo de cultivo.

Se volvió a correr el análisis bajo el método de maximización del R^2 y minimización del AIC, utilizando regresoras menores a 250 días de desarrollo del cultivo. El modelo que maximizó el R^2 y minimizó el AIC fue generado con las variables regresoras macollamiento, elongación 1, el acumulado (suma) de iniciación, macollamiento y elongación 1. Se obtuvo el modelo:

$$Productividad = 271.9(macollamiento) + 410.3(elongación1) - 117.3(acuinmacelo1) - 135.28$$

Donde: macollamiento es igual al NDVI promedio del día 46 al día 145, elongación 1 es igual al promedio del NDVI del día 146 al día 215, acuinmacelo1 es igual al promedio del NDVI de iniciación, macollamiento y elongación 1 sumados.

Comparado con el mejor modelo del primer grupo de modelos de regresión, el segundo modelo, presentó un menor índice de Akaike, o la máxima verosimilitud (1236.59), comparado con el primero que presentó un índice de Akaike de 1374. También mejoró el grado de ajuste de las variables regresoras a la variabilidad de la productividad R^2=0.75. Además, el segundo modelo cumple con la Ley de Parsimonia, que establece que la solución más simple suele ser la mejor, en este caso, el segundo modelo logró reducir el número de variables

regresoras a 3 variables y permite obtener un estimado de productividad a los 7 meses de desarrollo del cultivo.

Nuevamente la constante y la variable acumulada (acuinmael1) restan a la variabilidad de la productividad. Las variables macollamiento y elongación 1 suman a la variabilidad de la productividad.

El segundo modelo mostró que nuevamente la variable que más explica la variabilidad de la productividad fue la etapa de elongación 1, pero en el segundo modelo esta variabilidad explicada fue mayor 41.16 %, contra 12.9 del primer modelo. Se observa también que la variabilidad inherente al error se redujo en el segundo modelo a 24.7 %.

4.6. Validación de los supuestos de la regresión en el modelo de mejor ajuste

Se validaron los 3 supuestos principales del análisis de regresión: normalidad de los residuos, heterocedasticidad e independencia de varianzas. El *QQplot* de los residuos mostró que los datos siguieron una distribución parecida a la normal, por medio de la prueba de bondad de ajuste Kolmogorov Smirnov se determinó con un 95 % de confianza que efectivamente los residuos se distribuyen normalmente.

El supuesto más importante que los residuos deben cumplir para que el modelo sea válido es la homogeneidad de varianzas. Dado que los valores del estadístico de *Bresch-Pagan* no fueron significativos, se puede concluir que las varianzas de los residuos son iguales y que el modelo es adecuado para explicar la variabilidad de la productividad.

CONCLUSIONES

1. Se estableció en base al coeficiente de correlación lineal de Pearson y el valor P, que las categorías de NDVI por etapa fenológica son las categorías potenciales como regresoras que explican la variabilidad de la productividad de caña de azúcar porque presentaron grado de asociación alta positiva y fueron significativos.

2. Con el método de selección secuencial hacia atrás, se seleccionaron las variables regresoras que contribuyen a un mayor grado de ajuste fueron macollamiento, elongación 1, el acumulado de iniciación, macollamiento y elongación 1. El análisis de varianza determinó que la regresora que más explica la variabilidad de la productividad fue la etapa de elongación 1, en un 41.16 %.

3. Se determinó que el modelo que maximizó el R^2 (0.75) y minimizó el AIC (1236.59) fue:

$$Productividad = 271.9(macollamiento) + 410.3(elongación1) - 117.3(acuinmacelo1) - 135.28$$

4. Se determinó que el mejor modelo cumplió con los 3 supuestos del análisis de regresión, según la prueba de bondad de ajuste Kolmogorov Smirnov los residuos se distribuyeron normalmente. Según la prueba Bresch-Pagan existió homogeneidad de varianzas. La prueba Durbin-Watson determinó autocorrelación de varianzas, indicando que el modelo es

adecuado para predecir la variabilidad de la productividad de caña de azúcar.

RECOMENDACIONES

1. Para próximos estudios de NDVI y productividad se sugiere al ingenio, evaluar covariables categóricas como variedad de caña, porcentaje de despoblación inicial para mejorar la significancia de las categorías potenciales

2. Para la mejora continua del proceso, se hace necesario seleccionar las variables regresoras, en función de la proporción de la serie de NDVI, es decir, a medida que la serie tenga más imágenes durante el ciclo productivo las variables regresoras podrían mejorar su significancia.

3. En base a los resultados, se requiere desarrollar modelos de pronóstico por cada lote productivo para continuar con la mejora de ajuste del modelo.

4. Se propone evaluar modelos dinámicos que permitan pronosticar la productividad de caña a partir de los 5 meses de edad y que vayan mejorando el ajuste conforme avanza el ciclo productivo.

REFERENCIAS

1. Aguilar, N., Contreras, C., Galindo, G. y Fortanelli, J. (2010). Índice Normalizado de Vegetación en caña de azúcar en la Huasteca Potosina Avances en Investigación Agropecuaria, 14(2),49-65 Universidad de Colima México.

2. Akaike, H. (1985). A new look at the statistical model identification, IEEE Transactions on Automatic Control, vol. 19, pp. 716–723.

3. Alfaro, E. (2015). Evaluación de las perspectivas climáticas trimestrales realizadas por los institutos meteorológicos de Centro América. Santo Domingo, República Dominicana: Foro del clima, CRRH.

4. Aroca, P., García, C. y González, J. (2015). Estadística descriptiva e Inferencial. Recuperado de http://www.researchgate.net/publication/275021043

5. Badii, M., Guillen, A., Lugo Serrato, O. y Aguilar Carnica, J. (2014). Correlación No-Paramétrica y su aplicación en las Investigaciones Científicas. International Journal of Good Conscience, 31-40.

6. Bappel, E., Bégué, A., Martiné, J., Pellegrino A. y Siegmund, B (2005). Assimilation of a biophysical parameter estimated by remote sensing using spot 4&5 data into a sugarcane yield forecasting model.

7. Bastidas, E. y Carbonell, J. (2006). Soil spectral characterization and mineralogy of the Cauca River Valley by visible and infrared (400 - 2,500 nm) spectroscopy. Agronomía Colombiana. Recuperado de https://www.researchgate.net/publication/260769245_Soil_spectral_characterization_and_mineralogy_of_the_Cauca_River_Valley_by_visible_and_infrared_400_-_2500_nm_spectroscopy

8. Bégué, A., Baillarin, F., Bappel, E., Lebourgeois, V., Pellegrino, A. y Todoroff, P. (2010). Spatio-temporal variability of sugarcane fields and recommendations for yield forecast using NDVI, International Journal of Remote Sensing. http://dx.doi.org/10.1080/01431160903349057

9. Behar, R. (2003). Validación de supuestos en el modelo de regresión. Universidad del Valle, Cali: Serie Monografías

10. CENGICAÑA (Centro Guatemalteco de Investigación y Capacitación de la Caña de Azúcar). (2012). *El Cultivo de la Caña de Azúcar en Guatemala.* Melgar, M.; Meneses, A.; Orozco, H.; Pérez, O.; y Espinosa, R. (eds.). Guatemala. 512 p.

11. CENGICAÑA (Centro Guatemalteco de Investigación y Capacitación de la Caña de Azúcar). (2016). *Memoria. Presentación de resultados de investigación. Zafra 2015-2016*. Guatemala. En discos compactos -462 p. www.cengicana.org página 364.

12. FAO (Organización de las Naciones Unidas para la Alimentación y la Agricultura). (2015). Entendiendo el impacto de sequía provocada por El Niño en el área agrícola mundial: una evaluación utilizando

el Índice de Estrés Agrícola de la FAO (ASI). Serie sobre el medio ambiente y la gestión de los recursos naturales, 23(1), 1-42. Recuperado de http://www.fao.org/publications)

13. Fernandes, J., Favilla, N. y Dalla, J. (2017). Sugarcane yield prediction in Brazil using NDVI time series and neural networks ensemble, International Journal of Remote Sensing, 38(16), 4631-4644, DOI:10.1080/01431161.2017.1325531

14. Garcia, M. y Ortíz, A. (2017). Una nueva prueba para el problema de igualdad de varianzas (tesis de grado). Universidad Santo Tomás, Colombia.

15. Jiménez, E. (2004). *Introducción al análisis multivariable.* Recuperado de http://es.slideshare.net/tecnomexico/analisis-multivariable

16. Lofton, J., Tubana, B., Kanke, Y., Teboh, J., Viator, H. y Dalen, M. (2012). Estimating sugarcane yield potential using an in-season determination of normalized difference vegetative index, *Sensors*. 2012; 12(6):7529-7547. DOI.org/10.3390/s120607529

17. López, E. y González, B. (2015*). Estadística: Fundamentos y aplicaciones a la agronomía y ciencias afines*. Guatemala: Facultad de Agronomía, Universidad de San Carlos de Guatemala.

18. Martín, R. (S.F.) *Practicas estadísticas: Correlaciones con SPSS.* España: Escuela Superior de Informática, Universidad de Castilla La Mancha.

19. Martínez, D., Albín, J., Cabaleiro, J., Pena, T., Rivera, F. y Blanco, V. (2009). El Criterio de Información de Akaike en la Obtención de Modelos Estadísticos de Rendimiento. ResearchGate, (24), 439-444. Recuperado de htpp:// www.researchgate.net/publication/236279245.

20. Martínez, R., Rivadeneira, A. y Nieto, J. (2011). Guía de buenas prácticas para la predicción estacional en Latinoamérica. Perú: CIIFEN.

21. Mauricio, J. (2007*). Introducción al análisis de series temporales.* Madrid: Universidad Complutense de Madrid.

22. NCSS. (2015). Stepwise Regression. En A.A. NCSS. (Ed.), Stepwise Regression (pp. 1-9). Texas, Estados Unidos: NCSS Statistical software.

23. Rahman, M. y Robson, A. (2016). A Novel Approach for Sugarcane Yield Prediction Using Landsat Time Series Imagery: A Case Study on Bundaberg Region. *Advances in Remote Sensing*, 5, 93-102. http://dx.doi.org/10.4236/ars.2016.52008

24. Rudorff, B.F.T. y Batista, G.T. (1990). “Yield estimation of sugarcane based on agrometeorological-spectral based models”. Remote Sensing of Environment, 33, 183-192.

25. Rueda F, Peñaranda L., Velásquez W. y Díaz S. (2015). Aplicación de una metodología de análisis de datos obtenidos por percepción

remota orientados a la estimación de la productividad de caña para panela al cuantificar el NDVI (índice de vegetación de diferencia normalizada). Corpoica Cienc Tecnol Agropecu. 16(1): 25-40

26. Salvador Figueras, M. (2017*). Introduccion al Análisis Multivariante*. Retrieved from Estadistica: http://www.5campus.com/leccion/anamul

27. Schmidt, E.J., Narciso, G., Frost, P. y Gers, C.J. (2000). Application of remote sensing technology in the South African sugar industry: a review of recent research findings'. Proceedings of the South African Sugar Technologists' Association, 74, 192-201.

28. Siegel, S. (1956). Nonparametric Statistics for the Behavioral Sciences. McGraw-Hill. USA.

29. Sigüeñas, M. (2015). *Pruebas de normalidad*. Perú: Universidad Nacional Agraria La Molina

30. Simões, M., Rocha, J. y Lamparelli, R. (2005). Spectral variables, growth analysis and yield of sugarcane. *Scientia Agricola (Piracicaba, Brasil),* 62(3), 199-207.

31. Subirós, J.F., Sánchez, A., y Esquivel, E. (2010). Metodología empleada para estimar la producción de caña en azucarera el viejo, guanacaste. Disponible en: https://www.laica.co.cr/biblioteca2/buscar.do

32. Universidad Autónoma Chapingo. (2015). Boletín técnico informativo Cosecha de caña de azúcar en estado verde. Recuperado de https://www.gob.mx/cms/uploads/attachment/file/114363/1._Boletin_Julio_2015.pdf

33. Universidad de Vigo. (2017). Modelos autocorrelados: Un caso particular de los modelos de regresión lineal generalizado. Recuperado de https://docplayer.es/36757041-Modelos-autocorrelados-un-caso-particular-de-los-modelos-de-regresion-lineal-generalizado.html

34. Virginia, R., y Wall, D. (2001). Principles of Ecosystem function. En Levin, S.A. (Ed.) Encyclopedia of Biodiversity, (pp.345-352). San Diego, USA. Academic Press.

35. Webster, A. (2001). Estadística aplicada a los negocios y la economía. Bogotá: Irwin McGraw-Hill.

36. Wilks, D. (2006). Statistical methods in the atmospheric sciences. United States; Second edition. Cornell University.