UNIVERSIDAD SAN FRANCISCO DE QUITO USFQ

Colegio de Ciencias e Ingenierías

Aprendizaje Automático Multirrespuesta en la Industria de las Conservas de Frutas: un Camino hacia la Estandarización del Proceso Productivo

Ángel Isaac Burgos Naranjo
Daniel Sebastián Vásquez Játiva

Ingeniería Industrial

Trabajo de fin de carrera presentado como requisito
para la obtención del título de
INGENIERO INDUSTRIAL

Quito, 22 de diciembre de 2020

UNIVERSIDAD SAN FRANCISCO DE QUITO USFQ

Colegio de Ciencias e Ingenierías

HOJA DE CALIFICACIÓN
DE TRABAJO DE FIN DE CARRERA

Aprendizaje Automático Multirrespuesta en la Industria de las Conservas de Frutas: un Camino hacia la Estandarización del Proceso Productivo

Ángel Isaac Burgos Naranjo

Daniel Sebastián Vásquez Játiva

Nombre del profesor, Título académico **Danny Orlando Navarrete Chávez, MSc**

Quito, 22 de diciembre de 2020

© DERECHOS DE AUTOR

Por medio del presente documento certifico que he leído todas las Políticas y Manuales de la Universidad San Francisco de Quito USFQ, incluyendo la Política de Propiedad Intelectual USFQ, y estoy de acuerdo con su contenido, por lo que los derechos de propiedad intelectual del presente trabajo quedan sujetos a lo dispuesto en esas Políticas.

Asimismo, autorizo a la USFQ para que realice la digitalización y publicación de este trabajo en el repositorio virtual, de conformidad a lo dispuesto en la Ley Orgánica de Educación Superior del Ecuador.

Nombres y apellidos: Ángel Isaac Burgos Naranjo

Código: 00131200

Cédula de identidad: 1723581805

Nombres y apellidos: Daniel Sebastián Vásquez Játiva

Código: 00129996

Cédula de identidad: 1722166533

Lugar y fecha: Quito, 22 de diciembre de 2020

ACLARACIÓN PARA PUBLICACIÓN

Nota: El presente trabajo, en su totalidad o cualquiera de sus partes, no debe ser considerado como una publicación, incluso a pesar de estar disponible sin restricciones a través de un repositorio institucional. Esta declaración se alinea con las prácticas y recomendaciones presentadas por el Committee on Publication Ethics COPE descritas por Barbour et al. (2017) Discussion document on best practice for issues around theses publishing, disponible en http://bit.ly/COPETheses.

UNPUBLISHED DOCUMENT

Note: The following capstone project is available through Universidad San Francisco de Quito USFQ institutional repository. Nonetheless, this project – in whole or in part – should not be considered a publication. This statement follows the recommendations presented by the Committee on Publication Ethics COPE described by Barbour et al. (2017) Discussion document on best practice for issues around theses publishing available on http://bit.ly/COPETheses.

RESUMEN

La estandarización en la formulación de los alimentos es un desafío retador que comúnmente sufre el sector alimenticio, y la industria de las confituras, jaleas y mermeladas no es la excepción. La compañía ABC, firma dedicada a la producción y comercialización de productos agropecuarios, registra problemáticas en lograr niveles de consistencia, acidez y azúcares uniformes para sus cinco distintos sabores de mermeladas. La variabilidad propia de la materia prima, la estacionalidad y el estado de madurez de las frutas, inciden directamente en la calidad del producto terminado. El presente trabajo tiene como guía el uso de modelos estadísticos multivariados y de aprendizaje automático multirrespuesta como un camino para estudiar y modelar las relaciones entre las distintas variables de entrada y de salida que gobiernan la formulación de éste tipo de conservas. Se hallaron rangos de valores para los niveles de azúcar y acidez en las formulaciones que logran satisfacer matemática y estadísticamente los parámetros de liberación de producto terminado definidos por la compañía.

Palabras clave: mermeladas, modelos, variables, estándares, grados Brix, pH, consistencia.

ABSTRACT

Standardization in food formulation is a great challenge commonly faced by the food sector, and the jams, jellies and marmalades industry is no exception. ABC, a company dedicated to the production and commercialization of agricultural products, is experiencing problems in achieving uniform consistency, acidity and sugar levels for its five different jam flavors. The variability of the raw materials, the seasonality and the fruits' state of maturity, directly affect the quality of the finished product. The present work is guided by the use of multivariate statistical models and automatic multi-response machine learning algorithms as a way to study and model the relationships between the different input and output variables that govern the formulation of this type of preserves. Ranges of values for sugar and acidity levels in the formulations were found, which manage to mathematically and statistically satisfy the finished product quality release parameters defined by the company.

Key words: jams, models, variables, standards, Brix degrees, pH, consistency

TABLA DE CONTENIDO

ÍNDICE DE TABLAS

ÍNDICE DE FIGURAS

INTRODUCCIÓN

Generalidades

Las confituras, jaleas y mermeladas son productos alimenticios de consistencia gelatinosa derivados de la cocción de frutas, azúcares, ácidos comestibles y pectina (Fuster, 2004). Según Baker et. al. (2005) se cree que surgieron como un primer esfuerzo por conservar las frutas para su consumo fuera de temporada. Sin embargo, su popularidad no prosperó sino hasta mediados del siglo XIX cuando el azúcar se convirtió en una materia prima económicamente asequible en Europa (Endress et. al., 2006).

La elaboración de confituras, jaleas y mermeladas generalmente se la lleva a cabo de forma artesanal o industrial. Se utilizan frutas frescas, congeladas, esterilizadas o químicamente conservadas (Coronado & Hilario, 2001). En lo que corresponde a la industria, tradicionalmente se conocen dos tipos de sistemas productivos dependiendo si el proceso de cocción es abierto, utilizando presión atmosférica, o al vacío, utilizando presión reducida (Moyls et. al., 1962).

El uso de un hervidor abierto es una práctica adoptada por muy pocas o pequeñas compañías como un intento de mantener ciertas tradiciones en la elaboración del producto terminado (Baker et. al., 2005). La cocción al vacío es generalmente la preferida debido a que se trabaja con temperaturas más bajas y tiempos de procesamiento más cortos (Moyls et. al., 1962). Esto hace que los productos mantengan sus distintas características organolépticas, así como su apariencia e integridad (Moyls et. al., 1962).

En cuanto al mercado mundial de confituras, jaleas y mermeladas, en el 2014 éste fue de USD 20,087 millones y registró un consumo promedio per cápita de USD 3.55 (Williams & Marshall Strategy, 2020). Para el 2024 se espera que su valor alcance los USD 33,012

millones y su consumo promedio per cápita crezca con una tasa anual compuesta del 2.61% (Williams & Marshall Strategy, 2020).

Esto se debe a que éste tipo de productos son considerados como un atractivo complemento nutricional, particularmente en los desayunos, las meriendas y los postres (Fuster, 2004). En el Ecuador, la actividad económica que aborda éste sector comercial es la *C130.16 Elaboración de Compotas, Mermeladas y Jaleas, Purés y otras Confituras de Frutas o Frutos Secos* (Instituto Nacional de Estadísticas y Censos, 2012).

Según datos de la Superintendencia de Compañías, Valores y Seguros (2020) dicha actividad económica a nivel nacional reportó una utilidad promedio del ejercicio de USD 2,120.074, con una tasa de crecimiento anual del 21% para el período 2013-2018. En el presente trabajo se estudiará el caso de ABC, compañía dedicada a la oferta de soluciones alimenticias, así como a la producción y comercialización de mermeladas.

Problema

Para una mermelada artesanal las diferencias en firmeza y contextura no son de gran importancia (Endress et. al., 2006). No obstante, cuando se trata de un producto comercial, estos y otros atributos son considerados como características críticas de la calidad por parte de los consumidores. Por lo tanto, son un reto para una industria que no solo debe lidiar con la variabilidad propia de la materia prima, sino también con la variación intrínseca de la operación en planta (Endress et. al., 2006).

Featherstone (2016) sostiene que uno de los problemas más frecuentes en la fabricación de mermeladas es la poca gelificación del producto terminado. Por su parte Devi et. al. (2015) argumentan que otros inconvenientes son la cristalización de los azúcares, la aparición de burbujas o espumantes, y el deterioro microbiano. Con la finalidad de superar estas dificultades, es necesario estudiar las formulaciones de las mermeladas considerando la

interacción entre sus ingredientes más importantes y las distintas actividades que se llevan a cabo durante la totalidad de su proceso productivo (Broomfield, 1996).

Se conoce que la fruta, según su estado de maduración, sufre múltiples cambios bioquímicos y fisiológicos en su composición (Jordano, 2000). Conforme la fruta madura su contenido de pectina se degrada, su nivel de ácido disminuye y su cantidad de azúcar aumenta (Huber, 2001). La gran inestabilidad de la fruta con la que se trabaja hace que la adición y la compensación de ingredientes como la pectina o el ácido sea un paso crítico para alcanzar la gelificación adecuada del producto terminado y una consistencia óptima (Downing, 1996).

En el caso de ABC, la problemática es que se cuenta con formulaciones fijas que no necesariamente se adaptan a las condiciones iniciales de materia prima para cada corrida de producción. Los valores considerados para la mezcla y la adición de los ingredientes no varían según la naturaleza de las entradas del proceso, sino que se mantienen constantes, dificultando así la estandarización en la calidad del producto terminado.

Por este motivo, y considerando la situación antes mencionada, el estudio y la aplicación de modelos y procedimientos analíticos son necesarios para que la compañía ABC cuente con una investigación que conlleve a una mejor toma de decisiones, y logre sentar un camino hacia la mejora continua de su proceso productivo para sus cinco sabores estelares de mermeladas: mora, frutilla, piña, guayaba y frutimora.

Objetivos

Objetivo general.

- Estudiar la variabilidad en la producción de una familia de productos terminados de mermeladas, a través del uso de procedimientos analíticos multivariados y modelos de aprendizaje automático multirrespuesta, con la

finalidad de lograr una mayor uniformidad en sus estándares de calidad y en los parámetros de liberación definidos por la compañía.

Objetivos específicos.

- Modelar la formulación de una familia de productos terminados de mermeladas en función de su materia prima y su proceso productivo.
- Evaluar el rendimiento de procedimientos analíticos multivariados y modelos de aprendizaje automático multirrespuesta para la explicación y predicción de las variables de interés.
- Estudiar la sensibilidad de los modelos obtenidos bajo distintos escenarios de las variables de materia prima considerando su impacto en los parámetros de liberación entregados por ABC.

Propuesta de solución

En los siguientes apartados, el presente estudio desarrollará la aplicación de distintos procedimientos analíticos multivariados y modelos multirrespuesta de aprendizaje automático, con el fin de abordar la problemática de ABC desde distintos frentes. Se realizará una revisión literaria sobre trabajos similares en los que se haya hecho uso de este tipo de estadística aplicada en la resolución de problemas y la formulación productos en la industria alimenticia.

Por medio de una metodología integral, que contempla desde la definición del problema hasta el análisis y la interpretación de los resultados, los modelos estadísticos que se desarrollen podrán ser considerados como un camino para entender y predecir el efecto de las las variables de materia prima y de proceso, en la estandarización de la familia de productos terminados de mermeladas para el caso puntual de ABC.

DESARROLLO DEL TEMA

Revisión literaria

Estadística aplicada en la ciencia de alimentos.

La estadística, esencialmente, es la rama de las matemáticas que se ocupa de la recolección, la presentación, el uso y el análisis de los datos para la toma de decisiones (Montgomery & Runger, 2007). En este sentido, su estudio y aplicación en la ciencia de los alimentos juega un papel fundamental, al tratarse de un sector que debe lidiar con una gran cantidad de datos observacionales y experimentales (Bower, 2013).

Según Granato, de Araujo Calado y Jarvis (2014) , distintos procedimientos estadísticos son comúnmente requeridos para describir y estudiar los factores que influencian las propiedades químicas, físicas, sensoriales y microbiológicas de los alimentos. Entre algunas de las aplicaciones estadísticas más importantes en esta industria, resaltan el uso de métodos descriptivos e inferenciales, modelos de regresión y correlación, el control estadístico de la calidad y el diseño de experimentos (Bower, 1995).

Gosset y Pearson fueron los precursores de la estadística aplicada en los alimentos cuando a inicios del siglo XX introdujeron el concepto de significancia estadística en la formulación de cervezas Guinness (Box, 1987) . Luego fue Ronald Fisher quien profundizó en este campo, cuando hizo uso del diseño experimental para estudiar la variabilidad de los cultivos agrícolas en Reino Unido, así como las propiedades sensoriales del té (Bower, 2013).

Desde entonces, son muchos los trabajos sobre alimentos en los que se resumen y se analizan los datos procedentes de fuentes observacionales, experimentales, sensoriales e incluso directamente de los consumidores, con el fin de analizar diferencias, explorar relaciones, y poder evaluar la significancia estadística de los distintos resultados obtenidos sobre una o más hipótesis de interés para la industria (Kirk, Sawyer, & Egan, 1999).

Procedimientos analíticos multivariados en la industria alimenticia.

Los procedimientos estadísticos multivariados son aquellos métodos donde múltiples respuestas o variables dependientes se analizan simultáneamente con el fin de considerar sus relaciones y entender el comportamiento real de cada una de ellas en presencia de las demás (Alkarkhi & Alqaraghuli, 2019).

Los objetivos más importantes del análisis multivariado son, entre otros, simplificar y reducir la dimensionalidad de los datos, así como describir y predecir sus relaciones (Dattalo, 2013). En la industria alimenticia, el uso de éste tipo de estudios no es nuevo. Por ejemplo, existen aplicaciones en el estudio de vinos, colorantes y harinas.

En trabajos como el de Arvanitoyannis, Katsota, Psarra, Soufleros, y Kallithraka (1999) se hizo uso del análisis múltiple de varianza (MANOVA), el análisis de componentes principales (PCA) y el análisis canónico de discriminantes (CDA), con el fin de estimar los efectos de la variabilidad de los componentes químicos de los vinos en sus distintas respuestas sensoriales.

Por otra parte, en estudios como el de Lachenmeier y Kessler (2008) se implementó la conocida curva multivariada de resolución (MCR) para poder diferenciar y cuantificar las muestras espectrales obtenidas de distintas mezclas de colorantes artificiales. Éste método resultó efectivo para poder separar y analizar debidamente los colores, medidos con un espectrofotómetro.

Finalmente, Alkarkhi y Alqaraghuli (2019) mencionan que se utilizó PCA para distinguir entre cuatro tipos de harinas de cáscara y pulpa, elaboradas con bananas verdes y maduras. Hicieron uso de variables como grados Brix, color, pH y aceite. Se logró determinar el mejor método fisicoquímico para diferenciar estos tipos de harina, e identificar las variables que contemplan la mayor variabilidad para el estudio.

Modelos o algoritmos de aprendizaje automático en la industria alimenticia.

Según Burkov (2019), el aprendizaje automático es una rama de la informática dedicada a la construcción de algoritmos para la resolución de problemas prácticos. Gracias al uso de bases de datos y modelos estadísticos, éste campo del conocimiento se ha convertido en un factor diferenciador en las ciencias, las finanzas y la industria, debido a su gran poder y adaptabilidad (Friedman, Hastie, & Tibshirani, 2008).

Los algoritmos de aprendizaje automático se clasifican en dos grandes grupos: supervisados y no supervisados (Burkov, 2019). La diferencia más importante es que los algoritmos supervisados estructuran los datos para realizar predicciones, mientras que los no supervisados buscan descubrir estructuras y lograr establecer las reglas que rigen esos datos (Conway & Myles, 2012).

De igual manera, los modelos de aprendizaje automático, según su aplicación, generalmente se utilizan para métodos de regresión o clasificación (Friedman, Hastie, & Tibshirani, 2008). Burkov (2019) establece que un modelo de clasificación tiene un número finito de clases en donde el problema puede ser asignado, mientras que los algoritmos de regresión tienen un número infinito de clases.

En la industria alimenticia, Jiménez, González, Bagur y Cuadros (2019) afirman que las publicaciones de éste tipo de modelos en el estudio de los alimentos han registrado un gran crecimiento desde el año 2010. Jiménez et. al. (2019) señala que existe un énfasis particular en la publicación de trabajos que utilizan modelos de ensamble o de conjunto, los cuales emplean múltiples algoritmos simultáneamente con el fin de lograr un mejor rendimiento predictivo.

El bosque aleatorio, o random forest (RF) por su nombre en inglés, es uno de los más utilizados. En artículos como el de Martínez, Moreno, Cazares y Winkler (2017), se hizo uso de RF, en conjunto con PCA, para clasificar el tequila tradicional del mezcal. De manera

similar, en trabajos como el de Teye, Huan, Han y Botchway (2014), se empleó modelos como el de máquinas de vectores de soporte (SVM) o el de los los k vecinos más cercanos (kNN) para clasificar distintas muestras de cocoa con base en su país de procedencia.

Finalmente, en un artículo publicado por Estellez-Lopez (2017) se aplicó distintos modelos de regresión para evaluar el deterioro físico, químico y biológico de la carne. El objetivo de este estudio fue predecir el deterioro de la carne según el conteo de bacterias presentes en ella.

Métodos analíticos para la estandarización de las mermeladas.

Uno de los métodos analíticos más comúnmente utilizados en la estandarización de las mermeladas es el diseño experimental. En un artículo por Inam, Hossain, Siddiqui y Esdani (2012), se hizo uso de esta técnica estadística, donde mediante un diseño completamente aleatorizado por bloques, se logró evaluar y estimar el efecto de diez distintas variables en la formulación de mermeladas de malta, mango y piña.

Similarmente, en otro estudio por Besbes, Drira, Christope, Deroanne y Attia (2009), se investigó las propiedades físicas y sensoriales de distintos tipos de mermeladas, mediante la realización de un análisis de varianza (ANOVA) factorial, con la intención de determinar la existencia de diferencias estadísticas significativas en las propiedades antes mencionadas, en cada una de las mermeladas.

En lo que concierne a Latinoamérica, en un trabajo reciente, Garrido, Lozano y Genovese (2015) estudiaron y modelaron el efecto de los ingredientes más importantes de una mermelada de manzana, en sus propiedades reológicas y mecánicas. La implementación de un diseño experimental compuesto central optimizó la formulación de dicha mermelada, y maximizó su aceptabilidad general en sus distintas características organolépticas.

El presente trabajo busca estudiar la formulación y estandarización de mermeladas desde otro frente distinto al diseño experimental, siendo éste el de la construcción de modelos estadísticos multivariados y de aprendizaje automático multirrespuesta. Como se lo vio en la literatura, estos son procedimientos frecuentemente utilizados en la ciencia de alimentos, por lo que su implementación es una decisión no solo acertada sino también innovadora, en beneficio de ABC.

Metodología

Con el objetivo de abordar la problemática de ABC de forma integral y estructurada, el presente proyecto considerará como metodología una adaptación del proceso de modelado estadístico descrito por Shmueli (2010) para la construcción de procedimientos analíticos multivariados y modelos de aprendizaje automático multirrespuesta.

Según Shmueli (2010), independientemente si los modelos son de carácter descriptivo, explicativo o predictivo existe una serie genérica de pasos a cumplir para que su realización sea satisfactoria. La adaptación metodológica de estas etapas se aprecia en la Figura 1, y se las detalla brevemente a continuación.

Figura 1. Pasos de la Metodología Adaptada

Definición del problema y los objetivos.

Se debe analizar la situación lo suficiente como para poder identificar el problema con precisión y entender con claridad todas las preguntas de investigación. Es necesario definir los objetivos del proyecto considerando cómo o en qué manera los involucrados o interesados

utilizarán los resultados. Lo que se busca es traducir el problema en símbolos matemáticos para desarrollar y resolver los distintos modelos estadísticos (Shmueli et. al., 2020).

Diseño del plan de recolección de datos.

El diseño del plan de recolección de datos, según Shmueli y Koppius (2006), debe considerar si los datos con los que se va a trabajar son observacionales o experimentales, y si es que son de fuentes primarias (datos recolectados para efectos del estudio) o secundarias (datos recolectados para otros propósitos). En el primer caso, además, se debe contemplar entre otros factores, el tamaño de muestra, el esquema de muestreo y los instrumentos de medición.

Preparación y preprocesamiento de datos.

En esta etapa se debe corroborar que los datos se encuentren en condiciones razonables para su utilización. Se debe decidir cómo manejar los datos faltantes, y analizar si es que los valores se encuentran en los rangos esperados para cada una de las variables (Shmueli, 2010). Se debe, además, garantizar la coherencia de los datos en la definición de campos, unidades de medida y períodos de tiempo; e incluso crear nuevas variables que puedan aportar información adicional a los modelos (Shmueli et. al., 2020).

Análisis exploratorio de datos.

El análisis exploratorio de datos es un paso clave tanto para modelos explicativos como predictivos. En él se resumen y analizan los datos tanto numérica como gráficamente. En este punto se busca capturar relaciones tanto entre los distintos predictores, como entre ellos y las respuestas de interés (Shmueli, 2010). Se debe corroborar, además, que los datos cumplan con los distintos supuestos necesarios de los modelos, y evaluar posibles transformaciones de las variables (Shmueli & Koppius, 2006).

Selección de variables, modelos o algoritmos.

Las variables de respuesta se eligen con base en el objetivo del estudio, los datos disponibles y la precisión de los sistemas de medición. Por otro lado, las variables independientes se escogen según la teoría, el conocimiento y la evidencia empírica de su grado de asociación con las salidas del modelo (Shmueli, 2010).

La selección de modelos y algoritmos también debe alinearse con los objetivos y las características específicas del problema (Shmueli et. al., 2020). Shmueli y Koppius (2006) indican que los tipos de modelos a escoger típicamente recaen en tres categorías: algoritmos basados en datos, algoritmos de reducción de dimensionalidades y algoritmos múltiples o de ensamble.

Evaluación y validación de modelos o algoritmos.

Es posible evaluar el poder explicativo o predictivo en la construcción de modelos estadísticos. En el primer caso se mide el grado de relación entre las entradas y las salidas del modelo, en tanto que en el segundo se estudia la exactitud en la predicción. En modelos explicativos se valida el ajuste de los datos al modelo, mientras que en modelos predictivos se valida su generalización para poder predecir nuevas observaciones (Shmueli, 2010).

Análisis e interpretación de los resultados.

El análisis e interpretación de los resultados es una de las actividades más importantes en el desarrollo de un proyecto. Es en éste punto donde se incluye las implicaciones, recomendaciones y conclusiones de los modelos propuestos como solución. Es necesario probar y validar los algoritmos con los involucrados en el proyecto, verificar que no existan errores, y corregirlos de ser el caso (Shmueli et. al., 2020).

Aplicación a ABC

Definición del problema y los objetivos.

La compañía ABC tiene problemáticas con la estandarización de sus cinco distintos sabores de mermeladas. Existen casos donde los lotes de envasado no gelifican correctamente y por ende reportan una consistencia inadecuada. Estos tipos de inconvenientes obligan a que los lotes deban ser sometidos a reprocesos para poder ser liberados y satisfacer así las expectativas de los consumidores.

La empresa utiliza una formulación fija para cada uno de sus sabores sin que esto necesariamente considere la variabilidad inherente de la materia prima. Dicha variabilidad, por lo general, es propia de la estacionalidad y el estado de madurez de las frutas. Factores como los detallados anteriormente suelen traducirse en la producción de lotes no conformes y sobrecostos relacionados a los distintos reprocesos necesarios para su liberación.

Considerando los motivos antes detallados, el objetivo general del presente estudio es analizar y modelar la variabilidad en la formulación de una familia de productos terminados de mermeladas, mediante el uso de modelos estadísticos multivariados y de aprendizaje automático multirrespuesta, con la finalidad de entender y describir las relaciones entre las múltiples variables de interés para la firma ABC.

Diseño del plan de recolección de datos.

En un inicio, para cada uno de los cinco distintos sabores de mermeladas se definieron recolectar tres variables dependientes y seis independientes. Estas fueron definidas en función de distintos factores, donde se priorizó su importancia para el estudio, así como su disponibilidad y accesibilidad para su recolección. Se contemplaron predictores tanto de la materia prima como del proceso productivo con el fin de analizar gran parte de las fuentes de

variabilidad de los lotes de envasado o producto terminado. Las variables propuestas inicialmente se muestran en la Tabla 1, a continuación.

Tabla 1. *Listado Inicial de Variables Dependientes e Independientes*

Variables Independientes		**Variables Dependientes**
Materia Prima	**Proceso**	**Respuestas**
Sólidos Solubles [°Bx]	Tiempo Recepción-Uso	Consistencia [cm/s]
Nivel de Acidez [pH]	Tiempo de Cocción	Sólidos Solubles [°Bx]
Pectina [%/100 g]	Tiempo de Enfriamiento	Nivel de Acidez [pH]

Luego de definir toda la información necesaria para el levantamiento de los datos se hizo uso del módulo MVSAMPSI en STATA, mismo que fue exclusivamente diseñado para el cálculo del tamaño de muestra y poder estadístico en modelos multivariados (Moore, 1998). Éste módulo tiene como argumentos al estadístico Wilk's Lambda, la cantidad de variables dependientes e independientes utilizadas, y el poder de prueba deseado.

De acuerdo con Todorov y Filtsmore (2010), el estadístico de prueba Wilk's lambda es frecuentemente reportado en distintos procedimientos analíticos multivariados. Su propósito radica en evaluar, en una escala del cero al uno, la contribución de las variables independientes y su importancia relativa en los modelos. Considerando dicha escala, se eligió un valor de 0.75 debido a que es considerado como un límite conservador, ampliamente utilizado para estos propósitos (Stevens, 2009).

Según Montgomery (2013), el poder estadístico o de prueba busca minimizar la ocurrencia de un error tipo II al momento de evaluar las hipótesis sobre la existencia o no de términos significativos en los modelos. En este contexto, el error tipo II o beta consiste en considerar algún factor como significativo, cuando realmente éste no tiene efecto estadístico alguno en el algoritmo. Se eligió un valor de 0.80 para el poder de prueba deseado, puesto que

permite reducir el tamaño de muestra necesario, mantener una confiabilidad del 95%, y alcanzar una probabilidad moderadamente baja de incurrir en el tipo de error antes señalado (Montgomery, 2013).

El cálculo del tamaño de muestra se exhibe en el Anexo A, donde se observa que se determinó un mínimo de 73 observaciones por sabor para satisfacer los argumentos requeridos. Sin embargo, debido a la limitación de recursos para llevar a cabo el muestreo de la información, se decidió hacer uso de los datos históricos disponibles para el proceso de producción de mermeladas, mismos que fueron entregados por la empresa. Estos datos son de fuentes secundarias, debido a que no fueron diseñados directamente para efectos del estudio.

El alcance de la información disponible fue de enero a septiembre del año 2020, cubriendo así ocho meses de operación. De las seis variables independientes que fueron propuestas en un inicio, únicamente se contaron con dos de aquellas relacionadas a las frutas o la materia prima: sólidos solubles (°Bx) y nivel de acidez (pH).

No fue posible recolectar variables de entrada relacionadas al proceso productivo. Sin embargo, las variables dependientes o de salida, no se vieron modificadas; estas refieren a sólidos solubles (°Bx), nivel de acidez (pH) y consistencia del producto terminado. Los grados brix son útiles para cuantificar el contenido de azúcar tanto en las frutas como las mermeladas, y se los mide con un refractómetro (MacGillivray & Graham, 1969).

El pH indica el grado de acidez o alcalinidad del producto, y se lo gradúa con un potenciómetro (Rodriguez, Ocampo, & Escobar, 2012). La consistencia, por su parte, es una medida del grado de cohesión de las partículas que constituyen las mermeladas, y se la obtiene gracias a un consistómetro (Jordano, 2000).

Se debe destacar que los datos disponibles reportaron una dimensionalidad de 574 registros o filas y 22 campos o columnas. Estas variables se resumen en el Anexo B, donde es

posible notar la presencia de predictores tanto cuantitativos como cualitativos, así como variables que únicamente cumplen con un rol informativo o referencial (FECHA, LOTE, etc).

Análisis exploratorio de datos.

En primer lugar se normalizaron los datos disponibles con el fin de contar con una observación por registro, una variable por columna y una sola unidad de información por tabla (Wickham, 2014). Éste proceso facilitó la generación de una base de datos independiente para cada uno de los cinco sabores de mermeladas. Luego de esto se evaluó, por cada sabor, la integridad y la redundancia de la información en cuanto a la presencia de valores nulos y registros duplicados.

Según Shmueli et. al. (2020), los primeros tipos de valores son aquellos datos desconocidos o faltantes, en tanto que los segundos corresponden a observaciones consecutivas con valores idénticos en todos sus campos o columnas. Similarmente, también se estudió la presencia de registros catalogados como reprocesos, los cuales hacen referencia a la recolección de distintas observaciones para corridas de producción de un mismo lote de semielaborado.

Debido a que la información proporcionada por los duplicados y los reprocesos afectan directamente las operaciones matemáticas y los modelos estadísticos, fueron removidos. Los valores nulos fueron rellenados con la mediana de sus respectivas variables, puesto que según Shmueli et. al. (2020) este criterio de imputación impacta mínimamente en la distribución original de los datos, y mantiene en gran medida sus propiedades tanto de tendencia central como de dispersión.

Se debe hacer énfasis en que los datos atípicos, por su parte, fueron debidamente identificados y consensuados en conjunto con la compañía para poder ser retirados. En los Anexos C, D, E se presentan la cantidad de datos nulos, mediante diagramas de barras, así

como las gráficas de las distribuciones originales por cada sabor, mediante diagramas de caja y bigote. También se muestran sus respectivas caracterizaciones estadísticas.

En los Anexos F, G, H, por su parte, es posible apreciar las distribuciones una vez que se eliminaron y rellenaron los distintos tipos de datos que se destacaron anteriormente. Esto se resume en la tabla del Anexo I. La importancia de caracterizar estadísticamente la información obtenida radica en constatar que los valores de los factores de interés se encuentren en los rangos esperados para cada uno de los predictores (Shmueli, 2010).

Entre las variables continuas o numéricas, tanto de entrada como de salida, se calcularon correlaciones estadísticas de Pearson con la finalidad de identificar y cuantificar el grado de asociación lineal, positiva o negativa, entre los distintos factores (Anexo J). Por otro lado, en el caso de las variables cualitativas, para la variable de ESTADO o STATUS, la cual corresponde al estado de aprobación del lote de envasado, se realizaron diagramas de caja y bigote con el objetivo de identificar, gráficamente, posibles causas raíz de que un lote necesite ser reprocesado (Anexo K).

Finalmente, se evaluaron los supuestos de normalidad en las variables de salida. Éste es requisito indispensable para la validez de técnicas paramétricas de análisis como los modelos estadísticos multivariados. Si no se cumple con el supuesto de normalidad al emplear dichas técnicas, las pruebas de hipótesis estadísticas pueden verse afectadas (Statistics Solutions, 2020). Mediante el uso de una prueba Shapiro-Wilk (SW), se identificó aquellas variables de salida con distribuciones distintas a la normal (Anexo L).

El test SW plantea la hipótesis nula que una muestra proviene de una distribución normal, bajo una confiabilidad del 95%. Si bien existe una gran variedad de pruebas de normalidad, se utilizó la Shapiro-Wilk debido a que presenta una robustez considerablemente moderada hasta un tamaño de muestra de 2000 observaciones, siendo éste el caso que se tiene para cada uno de los cinco sabores de mermeladas (Royston, 1982).

Se buscó transformar aquellas variables no normales implementando el método de escalera de potencias de Tukey, donde con la ayuda del estadístico chi-cuadrado más bajo se eligió la transformación más óptima en cada caso (Anexo M). Sin embargo, cuando se corrió nuevamente la prueba SW, posterior a la transformación, se determinó que ciertas variables no lograron ser transformadas con éxito (Anexo N).

Selección de variables, modelos o algoritmos.

En relación al conjunto de datos que fue entregado por la compañía (ANEXO B), se consideraron únicamente los campos relacionados a brix (BRIX), pH (PH) y consistencia (CONSISTENCIA) del producto terminado, y a brix (BRIX_MMPP) y pH (pH_MMPP) de la materia prima. Esta decisión se la tomó debido a que, como se lo mencionó en la introducción del presente trabajo, son potenciales precursores de las respuestas de interés (Bower, 1995).

Se debe resaltar que, en el caso particular de la guayaba, dado que la materia prima es la pulpa y no la fruta, se contó con una variable independiente adicional, siendo ésta la consistencia, y a la que se la denominó CONST para efectos del estudio. Los factores asociados a las presiones de vacío y los porcentajes de acidez no fueron incluidos dado que tenían, como mínimo, un 80% de datos faltantes y por ello muy poca información relevante (Anexo C).

Los demás campos fueron descartados para su inclusión en los modelos por ser de carácter cualitativo o simplemente informativo, y con poco poder de predicción sobre las salidas. En cuanto a los tipos de algoritmos utilizados, estos fueron clasificados en tres grandes categorías: modelos estadísticos multivariados, modelos lineales generalizados (GLM) y modelos de aprendizaje automático multirrespuesta.

Los modelos multivariados fueron elegidos debido a la presencia de correlaciones moderadas entre las variables de respuesta (Anexo J), puesto que estos evalúan las pruebas de hipótesis sobre la existencia de términos significativos en los algoritmos considerando dichas

asociaciones. Más aún, de acuerdo con los autores Campbell y Friske (1959), el uso de procedimientos multivariados facilita el tener una visión integral del problema en cuestión.

En este sentido, concretamente, se decidió por hacer uso de modelos de regresión multivariada múltiple (MMR) debido a ciertas ventajas sobre otros procedimientos analíticos multivariados, como lo es su eficiencia computacional y facilidad de interpretación (Dattalo, 2013).

Para el caso de las variables de respuesta que no lograron ser transformadas a una distribución normal (Anexo N), se decidió emplear modelos lineales generalizados (GLM), los cuales son capaces de modelar distribuciones distintas a la normal, como las de Poisson, gamma, binomial, entre otras. Esto es algo que no es posible con las técnicas paramétricas convencionales (McCullagh & Nelder, 1989).

Finalmente, se llevaron a cabo modelos de aprendizaje automático multirrespuesta como una poderosa herramienta predictiva y de gran utilidad para la compañía. Fueron dos los tipos de modelos seleccionados por su gran adaptabilidad a pequeños conjuntos de datos, según Prakash (2018). El primero fue el modelo de árboles de decisión (o Decision Tree), mientras que el segundo fue el algoritmo k vecinos más cercanos (o k-Nearest Neighbors).

Formulación, evaluación y validación de modelos o algoritmos.

Previo a la implementación de los distintos tipos de modelos, se crearon predictores adicionales con la finalidad de capturar las relaciones no lineales entre las dos variables independientes disponibles. Esto se lo hizo cuidando que los términos añadidos no resulten en un posible sobreajuste de los modelos.

De acuerdo con Kenkel (2016), para modelos de regresión de mínimos cuadros parciales (OLS) de efectos fijos, como la regresión multivariada múltiple, no es necesario que

los modelos sean lineales en las covariables; estos mantienen sus propiedades siempre y cuando sean lineales en los parámetros. Las nuevas variables creadas se aprecian en en Anexo O.

Regresión multivariada múltiple.

Por cada sabor, así como por cada variable de respuesta normal, sea esta original o transformada, se corrió un algoritmo de regresión multivariada múltiple (Anexo P). Se determinó aquellos modelos estadísticamente significativos, y con la ayuda de un análisis múltiple de varianza (MANOVA), se identificó los factores estadísticamente significativos y sus efectos expresados en términos de coeficientes (Anexo Q).

Se evaluó el rendimiento de cada regresión obtenida con la ayuda del estadístico R2, el cual mide el porcentaje de variabilidad en los datos que explican los modelos (Montgomery, 2013). De igual manera, por cada algoritmo, se procedió a llevar a cabo un análisis de residuales (Anexo R) con el objetivo de corroborar si los resultados alcanzados eran estadísticamente confiables (Diaz, 2007).

En el caso de piña, para el modelo de consistencia (CONSISTENCIA), el análisis de residuales revela un sesgo considerable, y una distribución poco normal en las colas (Anexo R). Por este motivo se descartó su validez y posterior aplicación. Para los demás sabores, en todas sus variables de respuesta normales, no hubo mayor indicio en el ajuste y la distribución de los residuales que invaliden los resultados (Anexo R). Los modelos mutlivariados obtenidos se resumen en la Tabla 2, a continuación:

Tabla 2. *Modelos de Regresión Multivariada Múltiple*

Sabor	**Salida**	**Modelo MMR**	$\boldsymbol{R^2}$
Frutilla	pH	$\boldsymbol{pH} = 0.59ph + 1.40$	0.25
	Consistencia	No significativo	-
Frutimora	pH	$\boldsymbol{pH} = -0.25bx_{mora} + 0.015bx_{mora}^2 + 4.11$	0.08
	Consistencia	No significativo	-

Mora	pH	$\boldsymbol{pH} = -11.7ph + 42.9bx + 4.0ph^2 - 5.48bx^2 - 26.4(bx)(ph) + 3.9(ph^2)(bx) + 3.55(ph)(bx^2) - 0.57(ph^2)(bx^2) + 3.00$	0.19
	Consistencia	No significativo	-
Piña	pH	$\boldsymbol{pH} = -5.44ph + 0.87ph^2 + 10.89$	0.28
	Brix	No significativo	-
	Consistencia	$\boldsymbol{CONST.} = 65.76ph - 9.54ph^2 - 98.47$	0.29
Guayaba	pH	$\boldsymbol{pH} = 0.35ph - 0.012bx + 2.27$	0.39
	Brix	$\boldsymbol{BRIX} = 1.37ph - 0.25bx + 66.14$	0.33

Modelos lineales generalizados.

Como se lo mencionó anteriormente, para las variables de respuesta con distribuciones distintas a la normal, aún después de haber sido transformadas, se corrieron modelos lineales generalizados. Los resultados de los modelos se observan en el Anexo S, y se los resumen en la Tabla 3, donde además se incluyen dos distintos indicadores de rendimiento.

Tabla 3. *Modelos Lineales Generalizados*

Sabor	Salida	Modelo GLM	AIC	Bondad de Ajuste
Frutilla	Brix	$\boldsymbol{BRIX} = 4.78ph + 49.13$	444.96	2.11
Frutimora	Brix	No significativo	-	-
Mora	Consistencia	No significativo	-	-
Guayaba	Consistencia	$\boldsymbol{CONST.} = 0.57bx - 2.25$	221.09	1.43

El primer indicador es el Criterio de Información de Akaike (AIC), el cual es útil para la comparación y el contraste entre algoritmos. Mientras más bajo es su valor, se tiene un mejor balance entre el ajuste y el sobreajuste de los modelos (Yamaoka, Nakagawa, & Uno, 1978). El segundo indicador corresponde la Regla de Bondad de Ajuste para este tipo de regresiones, propuesta por Montgomery, Peck y Vining (2015).

Un resultado mayor a uno en esta última prueba revela que el modelo presenta un ajuste inadecuado (Montgomery et. al., 2015). Los algoritmos que se exhiben en la Tabla 3 fueron aquellos que reportaron un menor valor de AIC, y además se observa que su ajuste es adecuado al no diferir considerablemente de la unidad.

Modelos de aprendizaje automático multirrespuesta.

Se corrieron modelos de aprendizaje automático multirrespuesta de árboles de decisión y k vecinos más cercanos como una herramienta de utilidad para poder predecir los resultados de nuevas o futuras observaciones en cada uno de los cinco sabores de mermeladas. Con la finalidad de que los modelos y algoritmos desarrollados sean generalizables a nuevos conjuntos de datos, se evaluó su rendimiento.

Esto se lo hizo, particularmente, mediante el cálculo de la métrica del error absoluto medio (MAE), el cual evalúa la precisión en las predicciones de un modelo comparando los datos predichos frente a los observados (Xu, et al., 2019). Esta métrica facilitó el determinar el tipo de modelo con mejor rendimiento para cada sabor.

Se priorizó el error medio absoluto sobre otras métricas como el error cuadrático medio o la raíz del error cuadrático medio, dado que, según expertos en la evaluación de rendimiento de modelos, como Willmcott y Matsuura (2005), éste facilita su interpretación inmediata al no modificar la naturaleza propia de los datos originales ni de los predichos.

Mientras menor es el error, mayor el poder predictivo. Las métricas se exhiben en el Anexo T. Se debe decir que para la validación de estos modelos, se particionó la base de datos y se usó un 75% de ellos para entrenar el modelo, y el 25% restante para llevar a cabo las predicciones.

En el Anexo U, se muestran los resultados de las predicciones de los valores reales versus los predichos para el caso de consistencia (CONSISTENCIA). En estos gráficos es posible apreciar que no existe un sobreajuste de los modelos, más sí se pueden observar que existen observaciones donde los datos predichos se alejan significativamente de los reales.

Esto se debe, en gran medida, a la poca cantidad de datos disponibles para entrenar los modelos. No obstante, como se lo vio en el Anexo T, según la métrica del MAE, los datos

predichos se desvían ligeramente de los reales, dado que, los valores del MAE son bajos, inferiores a uno.

Análisis e interpretación de los resultados.

El análisis y la interpretación de los resultados tiene dos grandes frentes: la discusión del rendimiento de los modelos, y la discusión de su aplicabilidad práctica. En el primer frente, es necesario argumentar sobre la importancia no solo de la significancia estadística de los algoritmos, sino también sobre el porcentaje de variabilidad de los datos que estos son capaces de explicar.

En este sentido, según Anderson (2016), es posible enfrentarse con cuatro distintos escenarios: un R2 bajo con un valor p bajo, un R2 alto con un valor p alto, un R2 alto con un valor p bajo, y un R2 alto con un valor p alto. De acuerdo con Anderson (2016) , no existe un asociación o relación establecida entre estos valores.

Según Falk y Miller (1992), los valores recomendados de R2 deben ser iguales o superiores a 0.10 para poder argumentar que la varianza explicada de un modelo en particular es adecuada. En contraste, autores como Cohen (1998) denominan valores superiores a 0.13 como moderados, y Chin (1998) definen valores desde 0.33 como moderados.

Se debe destacar que se utilizó un nivel de signficancia de 0.05. Según los resultados obtenidos para los modelos de regresión multivariada múltiple y los modelos lineales generalizados (Tabla 1 y Tabla 3), estos recaen en en el primer escenario. Los resultados se deben, en gran parte, a la pobre especificación de los algoritmos en términos de sus variables independientes.

El contar únicamente con dos variables de entrada ignora las fuentes de variabilidad explicadas por otros factores discutidos en la literatura. Factores que no lograron ser recolectados por motivos de disponibilidad de recursos, como el porcentaje de pectina, así

como otros relacionados con el proceso productivo, sin duda alguna serían de gran utilidad para mejorar el rendimiento de los modelos y su generalización a escenarios mucho más realistas y complejos como los que enfrenta la compañía a nivel de producción.

Para la discusión de la aplicabilidad práctica de los modelos, en primer lugar se graficó su comportamiento en función de distintos valores de sus variables de entrada (Anexo V). Posteriormente, se llevó a cabo un análisis de sensibilidad con la finalidad de identificar qué rangos de dichas variables satisfacen los parámetros de liberación de los lotes de envasado entregados por la empresa. Estos se muestran en la Tabla 4, a continuación.

Tabla 4. *Rangos Permisibles de Variables de Materia Prima*

Sabor	Salida	Rangos pH MMPP	Rangos Brix MMPP
Frutilla	pH	[3.22-3.90]	-
	Brix	[3.32-4.16]	-
Frutimora	pH	-	[8.34-12.26]
Guayaba	pH	[2.36-4.65]	-
	Consistencia	-	[4.82-10.09]
Piña	pH	[3.46-3.91]	-

Para el caso de frutilla se tienen dos rangos de pH de materia prima. El primero de ellos satisface los parámetros de liberación del pH de producto terminado, en tanto que el segundo corresponde a sus grados Brix. El intervalo permisible, por lo tanto, es aquel que cumple con ambas salidas; es decir, de 3.22 a 4.16.

En frutimora se debe puntualizar que el rango de grados Brix indicado en la Tabla 4 corresponde únicamente a aquellos presentes en la mora, más no en la frutilla. Éste intervalo, si se lo compara con el rango histórico registrado por ABC para el período enero-septiembre 2020, resulta ser ligeramente más amplio.

En cuanto al sabor de guayaba, éste es similar al de frutilla en el sentido de que se tienen dos distintas salidas: pH y consistencia (CONSISTENCIA) del producto terminado. En contraste con los rangos históricos reportados por ABC, los límites permisibles de PH de

materia prima obtenidos por el modelo son más amplios, en tanto que los intervalos de grados Brix son más cortos.

Finalmente, en piña, de manera similar, se puede concluir que el rango de pH de materia prima es menos amplio que el histórico. Es posible argumentar que las diferencias entre los valores permisibles teóricos determinados por los distintos tipos de modelos y los valores observados históricamente, se deben a que no todos los efectos y las interacciones propuestas por la literatura para los distintos sabores fueron modelados por falta de datos y predictores.

CONCLUSIONES

La metodología de modelado estadístico propuesta por Shmueli et. al. (2010) fue una base imprescindible para poder llevar a cabo un trabajo integral, sistemático y correctamente estructurado. Todo estudio relacionado con el análisis de datos tiene como primer paso su manejo, limpieza y preprocesamiento, y éste sin duda alguna no fue la excepción.

Es común que en la industria se utilicen conjuntos de datos redundantes, con valores nulos, duplicados y atípicos, por lo que su gestión oportuna fue clave para el desarrollo de este proyecto. Gracias a la información entregada por la compañía ABC se determinó que, para el período de enero a septiembre del 2020, se registró entre un 5% y un 20% de lotes de envasado que requirieron ser reprocesados.

Si bien son distintas las causas que ocasionan esta problemática, mediante el análisis exploratorio de datos se identificó que, en gran medida, se debe a la compleja variabilidad existente en la consistencia de los lotes de envasado. Se halló, mediante un estudio de las distribuciones de los datos, que valores muy altos de consistencia (> 7 cm/s) se traducen en un fluimiento indeseado en las mermeladas.

Luego de haber calculado correlaciones estadísticas de Pearson entre las distintas variables de interés, se llegó a la conclusión de que las respuestas correspondientes a pH, brix y consistencia de producto terminado tienen una asociación lineal moderada (0.20-0.60), sea esta positiva o negativa.

Este hallazgo derivó en la utilización de modelos estadísticos multivariados para lograr el objetivo de modelar la formulación de mermeladas en función de distintos factores de interés para la compañía. Para esto se partió del supuesto de independencia de los datos, y se evaluó el supuesto de normalidad mediante una prueba de Shapiro-Wilk. Solo la variable pH resultó ser normal en todos los sabores, lo que conlleva a concluir, y coincidir con la literatura, en que

una gran cantidad de variables aleatorias que tienen lugar en la naturaleza o en la práctica no siguen una distribución gaussiana, y por lo tanto deben ser transformadas o analizadas con técnicas que no requieran de este supuesto para su validez estadística.

El método de escaleras de potencias de Tukey, como mecanismo de transformación de variables, resultó ser de gran utilidad para éste estudio, pese a que no fue efectivo en todas las salidas. Como alternativa, se implementaron modelos lineales generalizados con el propósito de estudiar variables de respuesta no normales, y adicionalmente se llevaron a cabo modelos de aprendizaje automático multirrespuesta como herramientas de gran utilidad en el ámbito predictivo.

Es posible argumentar que se alcanzaron, parcialmente, los objetivos planteados en un inicio. Se modeló la formulación de mermeladas, se encontró los factores significativos en cada uno de los modelos, y se estimó su efecto o importancia relativa en términos de coeficientes o parámetros. Para las tres grandes categorías de modelos, adicionalmente, se evaluó su rendimiento mediante el cálculo de distintas métricas.

Los valores pequeños de R2 que fueron obtenidos, revelan que los modelos explican relativamente poca variación de los datos. No obstante, esto se debe a la pobre especificación de los mismos en términos de otras fuentes de variabilidad (predictores) que contribuyen a la descripción de las distintas salidas. El no poder variar los porcentajes de pectina en fórmula, así como el no contar con otros predictores y posibles precursores de salida, sin duda alguna afectó el desempeño de las distintas regresiones.

Pese a ello, gracias a un análisis de sensibilidad, estos modelos fueron de gran utilidad para determinar rangos de las variables de materia prima que satisfagan los parámetros de liberación de lotes de envasado, establecidos por la compañía ABC. Estos rangos obedecen a las relaciones que las regresiones obtenidas fueron capaces de capturar, y deben ser validados por la compañía en la obtención de lotes de semielaborado.

RECOMENDACIONES

Se recomienda a la compañía ABC realizar corridas piloto entre los rangos de valores encontrados para las variables de pH y grados brix de lote semielaborado. Esto se sugiere con la finalidad de identificar si es que los rangos establecidos en realidad cumplen con los parámetros de liberación permitidos, no solo en la teoría sino también en la práctica. Lo que se busca es, eventualmente, definir límites de especificación para los lotes de semielaborado.

En cuanto a la recolección de los datos, se recomienda a la compañía el empezar a levantar otros tipos de variables que ayuden a entender de mejor manera la variabilidad en la consistencia, los grados brix y el pH de los lotes de envasado. Si bien al contar con una formulación fija, el contenido de pectina permanece constante, es posible muestrear información relacionada al proceso productivo para comprender sus interacciones.

Se propone, adicionalmente, como medida de control en la línea del proceso, fortalecer las técnicas relacionadas al aseguramiento y el control estadístico de la calidad. El uso de diagramas de control para variables numéricas continuas, como es el caso las variables estudiadas, es clave para monitorear en tiempo real las causas asignables raíz de aquellas observaciones no conformes.

Cuando el proceso se encuentre bajo control estadístico, además, es necesario realizar un análisis de capacidad del proceso para determinar si éste es capaz de producir salidas dentro de los rangos de liberación permitidos. Finalmente, si bien el diseño experimental es una opción poco viable en el sentido práctico por todos los recursos que implica, se propone implementar un estudio de mediano plazo relacionado a la operación evolucionaria (Montgomery, 2013).

La operación evolucionaria es un método utilizado para la mejora continua y monitoreo de operaciones a gran escala. Está diseñada como un método de rutina de planta que puede fácilmente ser realizado por los operarios con asistencia mínima del equipo de investigación y

desarrollo. Esta consiste en la introducción sistemática de pequeños cambios en en los niveles de las variables de interés (Montgomery, 2013).

Los cambios de las variables son tan pequeños que no afectan la producción, pero son lo suficientemente grandes como para poder encontrar mejoras potenciales en el rendimiento y la estandarización de los procesos (Montgomery, 2013). Por estos motivos se sugiere a la compañía ABC considerar, dentro de un corto y mediano plazo, la realización de éste tipo de estudios.

La operación evolucionaria generalmente emplea un diseño 2k, donde se puede trabajar con k variables, aunque generalmente se emplean dos o tres. El diseño se maneja por ciclos donde cada ciclo se completa cuando se toma una observación en cada nivel. Se deberá tener una observación en los niveles establecidos y tomar nuevos niveles ligeramente superiores e inferiores a los antes utilizados (Montgomery, 2013).

Se propone a ABC que en cada ciclo de la operación evolucionaria lleve a cabo pequeños cambios o modificaciones en las variables independientes de pectina, azúcares y ácidos comestibles añadidos, y como variable de respuesta, recolecte en cada ciclo la consistencia de los lotes de envasado, dado que es la principal variable de interés.

LIMITACIONES

La principal limitación de éste estudio fue el no poder contar con todas las variables predictoras del plan de muestreo original. Como se lo remarcó en el desarrollo de éste trabajo, únicamente se consideraron las variables independientes de pH y brix de lotes semielaborados en relación a las seis que se propusieron inicialmente. No fue posible contemplar ninguna variable de proceso, ni el contenido natural de pectina de la materia prima.

Éste último es el principal precursor de la consistencia, una de las variables de respuesta de mayor interés en el estudio. Es lógico argumentar que esta limitante hizo que, en cuatro de los cinco sabores, los modelos no logren explicar estadísticamente dicha salida (no fueron significativos).

Otra limitación es que la calidad y la cantidad de los datos impidieron encontrar las interacciones y relaciones propuestas por la literatura, dado que claramente se indica que los grados brix y pH influyen en la consistencia y otras salidas del producto terminado, pero esto no necesariamente se reflejó en los modelos y se tradujo, en ciertos casos, en un porcentaje de variabilidad muy bajo explicado por los algoritmos.

La poca familiarización de primera mano con el proceso también fue una restricción que impidió un entendimiento holístico de las operaciones de la compañía ABC, y por lo tanto la recolección de los datos. Los detalles propios de la producción en planta no fueron adentrados.

Finalmente, se lleva la limitación de no haber podido correr un diseño experimental factorial en un ambiente cuasi controlado. Éste es el método preferido por antonomasia para estandarizar las formulaciones de alimentos cuando se deben considerar interacciones complejas entre las distintas variables bajo estudio.

REFERENCIAS BIBLIOGRÁFICAS

Alkarkhi, A. F., & Alqaraghuli, W. A. (2019). *Easy statistics for food science with R.* London: Academic Press.

Alkarkhi, A., & Alqaraghuli, W. (2019). *Easy Statistics for Food Science with R.* London: Academic Press.

Anderson, F. (2016, June 15). *What is the relationship between R-squared and p-value in a regression?* Retrieved from ResearchGate: https://www.researchgate.net/post/What_is_the_relationship_between_R-squared_and_p-value_in_a_regression

Arvanitoyannis, I., Katsota, M., Psarra, E., Soufleros, E., & Kallithraka, S. (1999). Application of quality control methods for assessing wine authenticity: Use of multivariate analysis (chemometrics). *Trends in Food Science & Technology*, 321±336.

Baker, R. A., Berry, N., Hui, Y. H., & Barrett, D. M. (2005). Fruit preserves and jams. In D. Barrett, L. Somogyi, & H. Ramaswamy, *Processing fruits: science and technology* (p. 113). Boca Raton: CRC Press LLC.

Basu, S., & Shivhare, U. (2010). Rheological, textural, micro-structural and sensory properties of mango jam. *Journal of Food Engineering*, 357–365.

Besbes, S., Drira, L., Christope, B., Deroanne, C., & Attia, H. (2009). Adding value to hard date (Phoenix dactylifera L.): Compositional, functional and sensory characteristics of date jam. *Food Chemestry*, 406-411.

Bower, J. A. (1995). Statistics for food science I. *Nutrition and Food Science*, 19-23.

Bower, J. A. (2013). *Statistical methods for food science: introductory procedures for the food practitioner.* New Jersey: John Wiley & Sons, Inc. .

Box, J. (1987). Guinness, Gosset, Fisher, and small samples. *Statistical Science*, 45-52.

Bro, R., Van der Berg, F., Thybo, A., Andersen, C., Jorgensen, B., & Andersen, H. (2002). Multivariate data analysis as a tool in advanced quality monitoring in the food productionchain. *Trends in Food Science & Technology*, 235–244.

Broomfield, R. W. (1996). The manufacture of preserves, flavourings and dried fruits. In D. Arthey, & P. R. Ashurst, *Fruit processing.* Boston: Springer.

Burkov, A. (2019). *The hundred page machine learning book.* Andriy Burkov.

Chin, W. (1998). The partial least squares approach for structural equation modeling. In G. Marcoulides, *Modern Methods for Business Research* (pp. 295-236). London: Lawrence Erlbaum Associates.

Cohen, J. (1998). *Statistical Power Analysis for the Behavioral Sciences.* New York: Lawrence Erlbaum Associates Publishers.

Conway, D., & Myles, J. (2012). *Machine Learning for Hackers: Case Studies and Algorithms to Get you Started.* Sebastopol: O'Reilly Media.

Coronado, M., & Hilario, R. (2001). *Elaboración de mermeladas: procesamiento de alimentos para pequeñas y micro empresas agroindustriales.* Lima: Centro de Educación, Investigación y Desarrollo, CIED.

Dattalo, P. (2013). *Analysis of multiple dependent variables.* New York: Oxford University Press.

Devi, M. P., Bhowmick, N., Bhanusree, M. R., & Ghosh, S. K. (2015). Preparation of value-added products thorugh preservation. In A. B. Sharangi, & S. Datta, *Value addition of horticultural crops: recent trends and future directions* (p. 30). Pundibari: Springer India.

Diaz, J. (2007, October 7). Retrieved from Universidad de Puerto Rico: http://math.uprag.edu/residuales1.pdf

Downing, D. (1996). *A complete course in canning: processing procedures for canned food products.* CTI Publications Inc.

Endress, H.-U., Mattes, F., & Norz, K. (2006). Pectins. In Y. H. Hui, *Handbook of food science, technology, and engineering* (Vol. 3, p. 15). Boca Raton: CRC Press.

Estellez-Lopez, L., Ropodi, A., Pavlids, D., Fotopoulou, J., Gkousari, C., Peyrodie, A., . . . Mohareb, F. (2017). An automated ranking platform for machine learning regression models for meat spoilage prediction using multi-spectral imaging and metabolic profiling. *Food Research International,* 206-215.

Falk, F., & Miller, N. (1992). *A Primer for Soft Modeling.* Ohio: The University of Arkon Press.

Featherstone, S. (2016). Jams, jellies and related products. In S. Featherstone, *A complete course in canning and related processes* (pp. 333-336). Woodhead Publishing.

Friedman, J., Hastie, T., & Tibshirani, R. (2008). *The Elements of Statistical Learning.* Springer.

Fuster, V. (2004). Mermeladas y confituras. En P. López, J. Boatella, & R. Codony, *Química y bioquímica de los alimentos II* (pág. 105). Barcelona: Edicions Universitat Barcelona.

Garre, A., Ruiz, M. C., & Hontoria, E. (2020). Application of Machine Learning to support production planning of a food industry in the context of waste generation under uncertainty. *Operations Research Perspectives*.

Garrido, J. I., Lozano, J. E., & Genovese, D. B. (2015). Effect of formulation variables on rheology, texture, colour, and acceptability of apple jelly: modelling and optimization. *Food Science and Technology*, 325-332.

Granato, D., de Araújo Calado, V. M., & Jarvis, B. (2014). Observations on the use of statistical methods in food science and technology. *Food Research International*, 137-149.

Huber, D. J., & Karakurt, Y. (2001). Pectin degradation in ripening and wounded fruits. Gainesville, Florida, United States of America.

Inam, A., Hossain, M., Siddiqui, A., & Easdani, M. (2012). Studies on the Development of Mixed Fruit Marmalade. *Enviromental Science and Natural Resources*, 315-322.

Instituto Nacional de Estadísticas y Censos. (2012). *Clasificación Nacional de Actividades Económicas.*

Jimenez, A., Gonzalez, A., Bagur, G., & Cuadros, L. (2019). Alternative data mining/machine learning methods for the analytical evaluation of food quality and authenticity – A review. *Food Research International*, 25-39.

Jordano, P. (2000). Fruits and frugivory. In M. Fenner, *Seeds: the ecology of regeneration in plant communities* (pp. 125-166). Wallingford: CABI Publ.

Kenkel, B. (2016, February 4). *Higher order terms*. Retrieved from Reintroduction to linear regression: http://bkenkel.com/psci8357/notes/04-higher-order.html

Kirk, R. S., Sawyer, R., & Egan, H. (1999). *Pearson's composition and analysis of foods.* Harlow: Longman Scientific and Technical.

Lachenmeier, D., & Kessler, W. (2008). Multivariate Curve Resolution of Spectrophotometric Data for the Determination of Artificial Food Colors. *Journal of agricultural and food chemestry*, 5463–5468.

MacGillivray, A., & Graham, W. (1969). Brix Determination. *Proceedings of The South African Sugar Technologists' Association* , 215-2018.

Martinez, S., Moreno, A., Cazares, D., & Winkler, R. (2017). Automated chemical fingerprinting of Mexican spirits derived from Agave (tequila and mezcal) using direct-injection electrospray ionisation (DIESI) and low-temperature plasma (LTP) mass spectrometry. *Analytical Methods*.

McCullagh, P., & Nelder, J. (1989). *Generalized Linear Models* . Boca Raton: Champman & Hall/CRC.

Montgomery, D. (2013). *Design and Analisis of Experiments.* New Jersey: John Wiley & Sons.

Montgomery, D. C., & Runger, G. C. (2007). *Applied statistics and probability for engineers.* New Jersey: John Wiley & Sons, Inc.

Montgomery, D., Peck, E., & Vining, G. (2015). *Introduction to Linear Regression Analysis.* New Jersey: John Wiley & Sons.

Moore, D. E. (1998). MVSAMPSI: Stata module to determine sample size and power for multivariate regression. *Statistical Software Components*.

Moyls, A. W., Strachan, C. C., & Atkinson, F. E. (1962). *Making jam commercially: principles, methods, equipment, formulas.* Ottawa: Canada Dept. of Agriculture.

Prakash, J. (2018, December 21). *Breaking the curse of small datasets in Machine Learning*. Retrieved from Towards Data Science: https://towardsdatascience.com/breaking-the-curse-of-small-datasets-in-machine-learning-part-1-36f28b0c044d

Rodriguez, A., Ocampo, M., & Escobar, E. (2012). Acondicionamiento del sensor de pH y temperatura para realizar titulaciones potenciométricas. *Scientia Et Technica*, 188-196.

Royston, P. (1982). An extension of Shapiro and Wilks' W test for normality to large samples. *Applied Statistics*, 115-124. Retrieved from Stata Web site.

Shmueli, G. (2010). To explain or to predict? *Statistica Science*, 289-310.

Shmueli, G., & Koppius, O. (2006). *Predictive analytics in information systems research.* Paphos: Conference on Information Systems and Technology.

Shmueli, G., Bruce, P., Gedeck, P., & Patel, N. (2020). *Data mining for business analytics.* Hoboken: John Wiley & Sons.

Stevens, J. P. (2009). *Applied multivariate statistics for the social sciences.* Cincinnati: Routledge.

Superintendencia de Compañías, Valores y Seguros . (2020, Octubre 5). *Portal de información: sector societario*. Retrieved from Estados financieros por ramo: https://reporteria.supercias.gob.ec/portal/cgi-bin/cognos.cgi?b_action=cognosViewer&ui.action=run&ui.object=%2fcontent%2ffolder%5b%40name%3d%27Reportes%27%5d%2ffolder%5b%40name%3d%27Estados%20Financieros%27%5d%2freport%5b%40name%3d%27Estados%20Financieros

Teye, E., Huang, X., Han, F., & Botchway, F. (2014). Discrimination of Cocoa Beans According to Geographical Origin by Electronic Tongue and Multivariate Algorithms. *Food Analisis Methods*, 360-365.

Todorov, V., & Filzmoser, P. (2010). Robust Statistic for the One-way MANOVA. *Computational Statistics & Data Analysis*, 37-48.

Wickham, H. (2014). Tidy data. *Journal of Statistical Software*, 1-24.

Williams & Marshall Strategy. (2020, Mayo 21). *Research and markets*. Retrieved from Global jams, jellies and marmalades market and the impact of Covid-19 in the medium term: https://www.researchandmarkets.com/reports/5001359/the-global-jams-jellies-and-marmalades-

market?utm_source=dynamic&utm_medium=BW&utm_code=3jq3gf&utm_campaign=1390677+-+Global+Jams%2c+Jellies+and+Marmalades+Market+and+the+Impact+of+COVID-19+in+the+Medium+

Wilmcott, C., & Matsuura, K. (2005). Advantages of the mean absolute error (MAE) over the root mean square error (RMSE) in assessing average model performance. *Climate Research*.

Xu, D., Shi, Y., Tsang, I., Ong, Y., Gong, C., & Shen, X. (2019). A Survey on Multi-output Learning. *IEEE Transactions on Neural Networks and Learning Systems*.

Yamaoka, K., Nakagawa, T., & Uno, T. (1978). Application of Akaike's Information Criterion (AIC) in the Evaluation of Linear Pharmacokinetic Equations. *Journal of Pharmacokinetics and Biopharmaceutics*,.

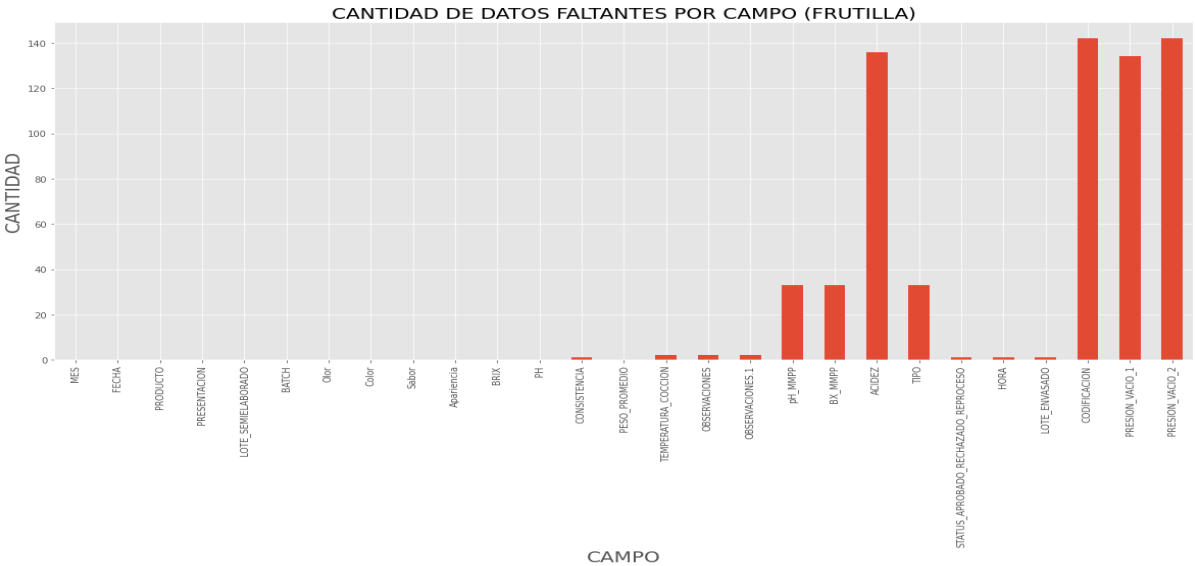
ANEXO A: CÁLCULO DE TAMAÑO DE MUESTRA

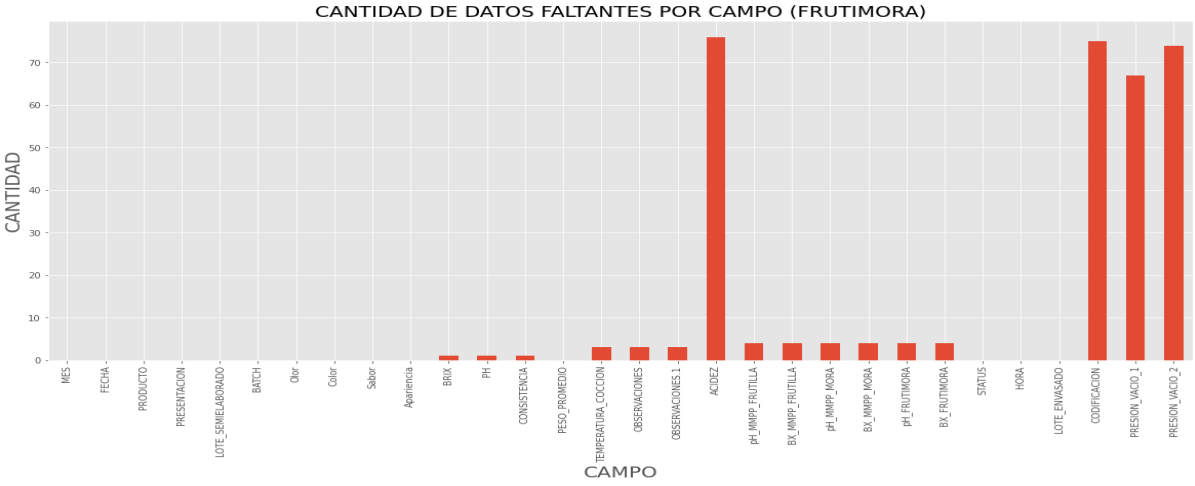
```
. mvsamp1i .75, ny(3) nx(6) p(.8)

            MULTIVARIATE POWER ANALYSIS

N.............................. 73
Alpha.......................... 0.0500
Power (Beta)................... 0.7946 (0.2054)
Wilks' Lambda.................. 0.7500
Effect Size.................... .1071
F.............................. 1.0796
Hypothesis df.................. 18
Error df....................... 181.5046
R-squared...................... 0.2500
Adjusted R-squared............. 0.0200
Noncentrality Parameter........ 21.4667
```

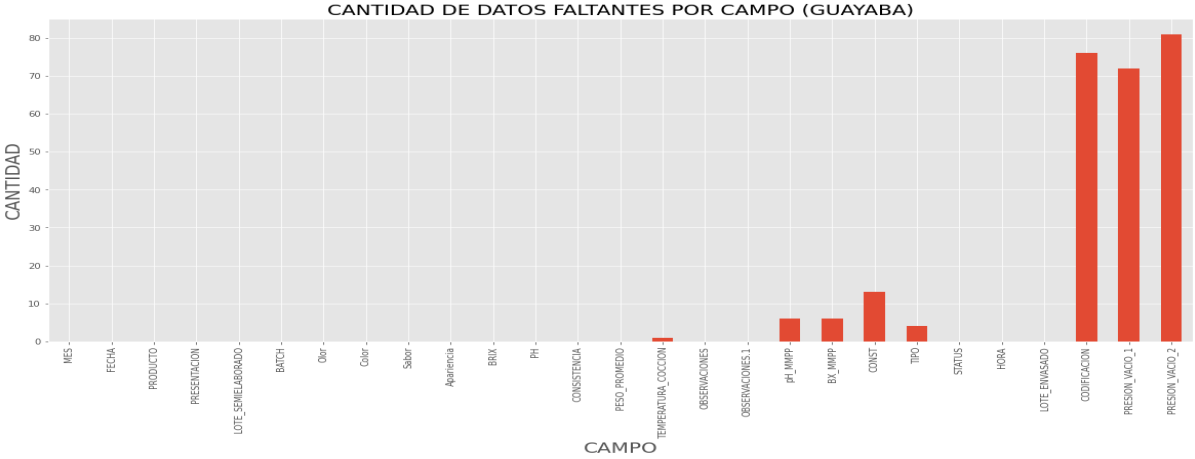
ANEXO B: VARIABLES ORIGINALES

VARIABLES ORIGINALES			
MES	BATCH	PESO PROMEDIO	LOTE ENVASADO
FECHA	OLOR	TEMPERATURA COCCION	CODIFICACION
PRODUCTO	COLOR	OBSERVACIONES	PRESION DE VACIO 1
PRESENTACION	SABOR	STATUS	PRESION DE VACIO 2
LOTE SEMIELABORADO	APARIENCIA	HORA	

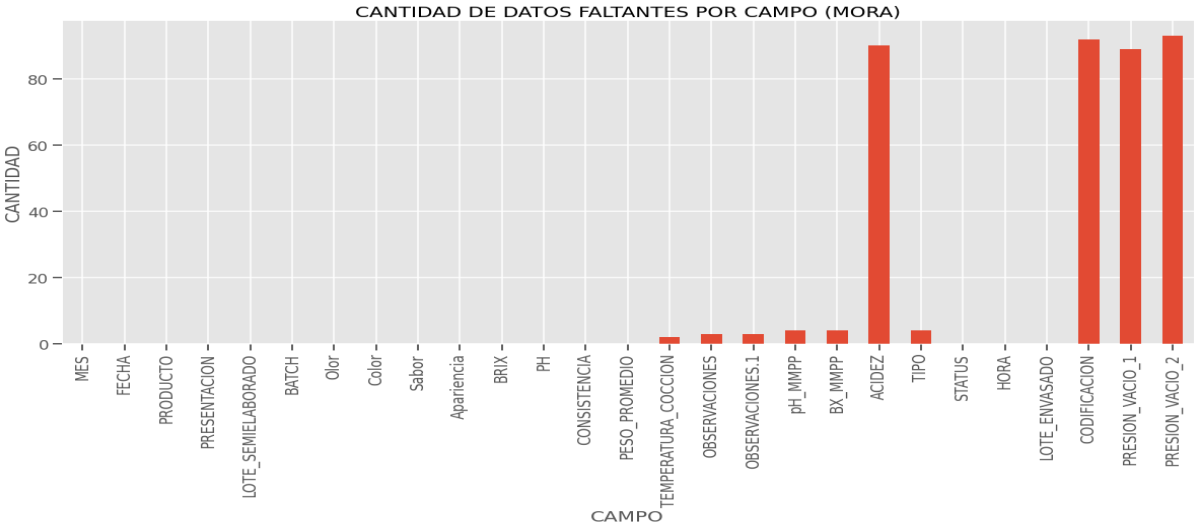
ANEXO C: CANTIDAD INICIAL DE DATOS FALTANTES

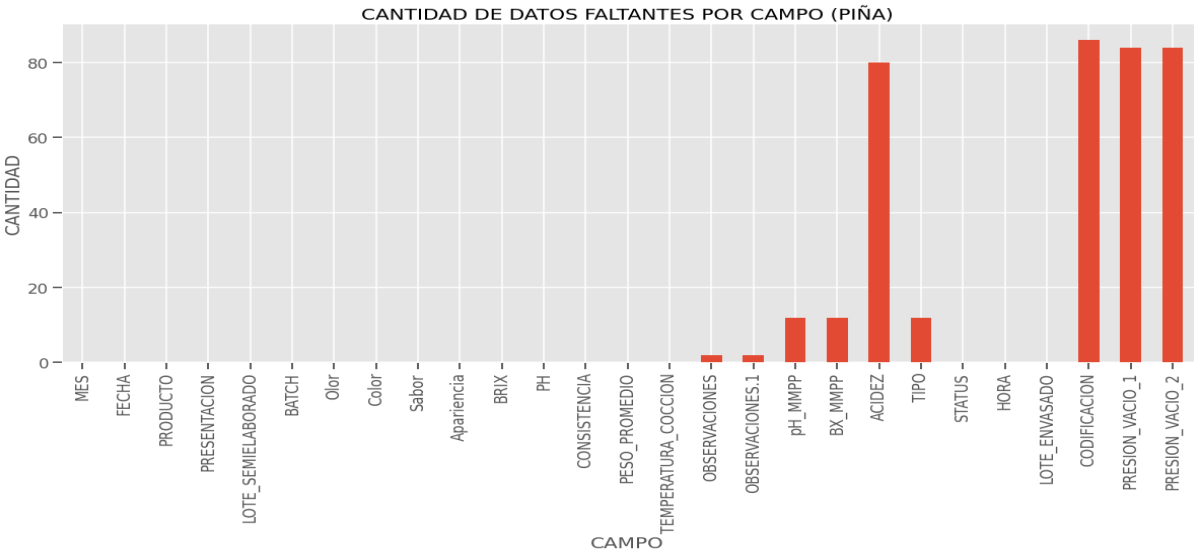




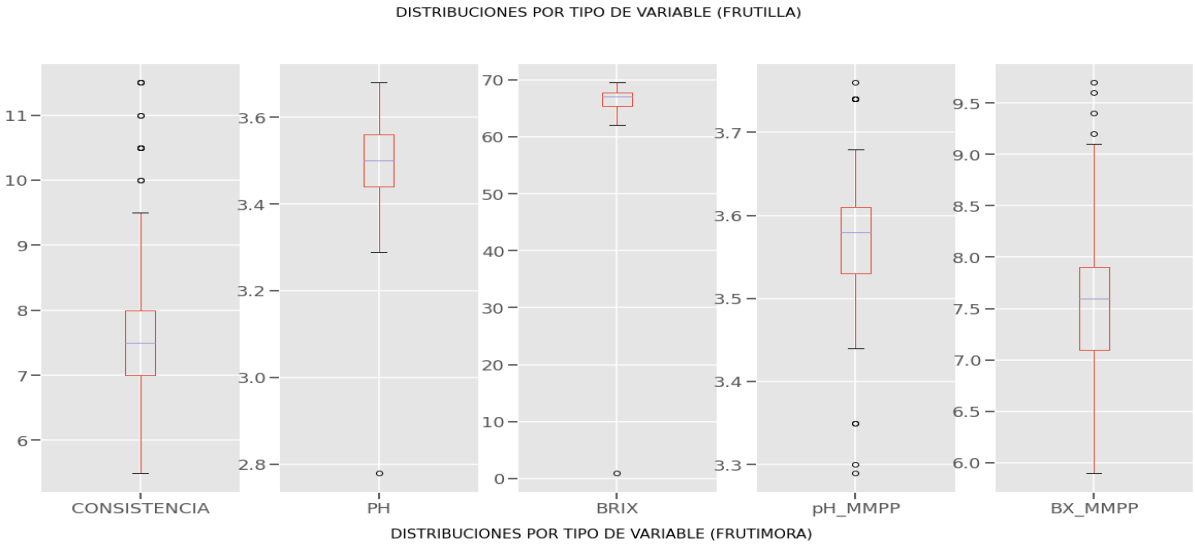


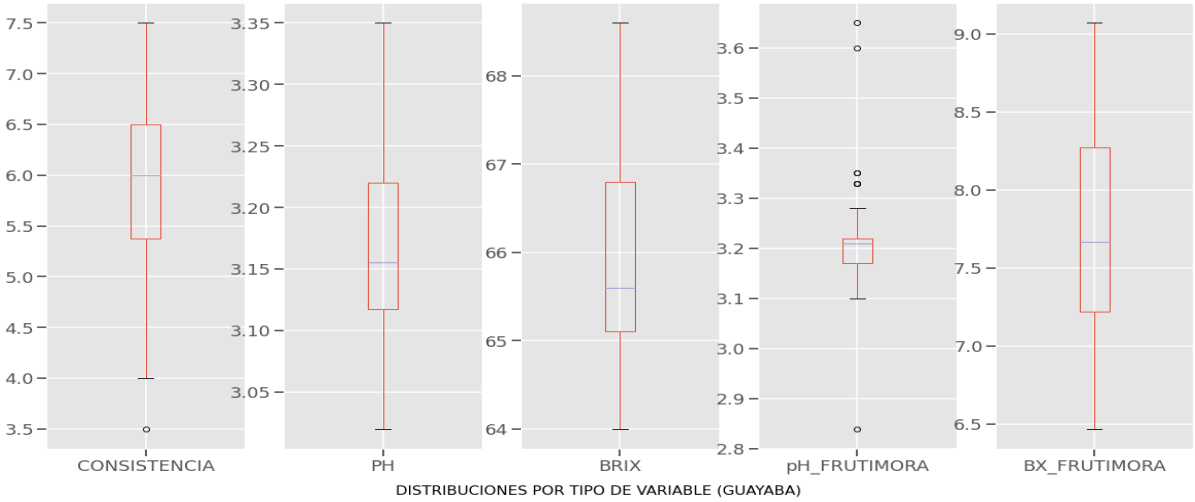
ANEXO C: CANTIDAD INICIAL DE DATOS FALTANTES (CONT.)

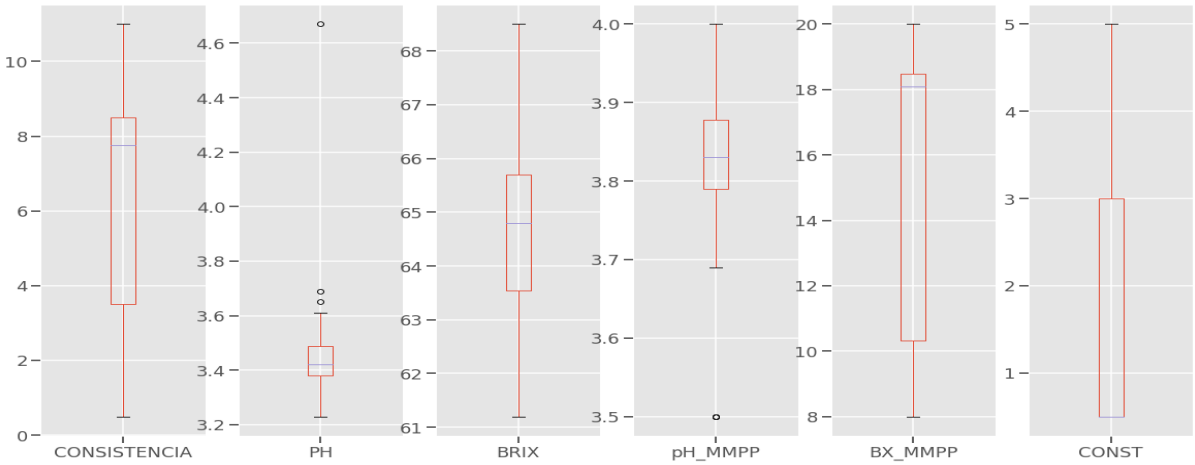




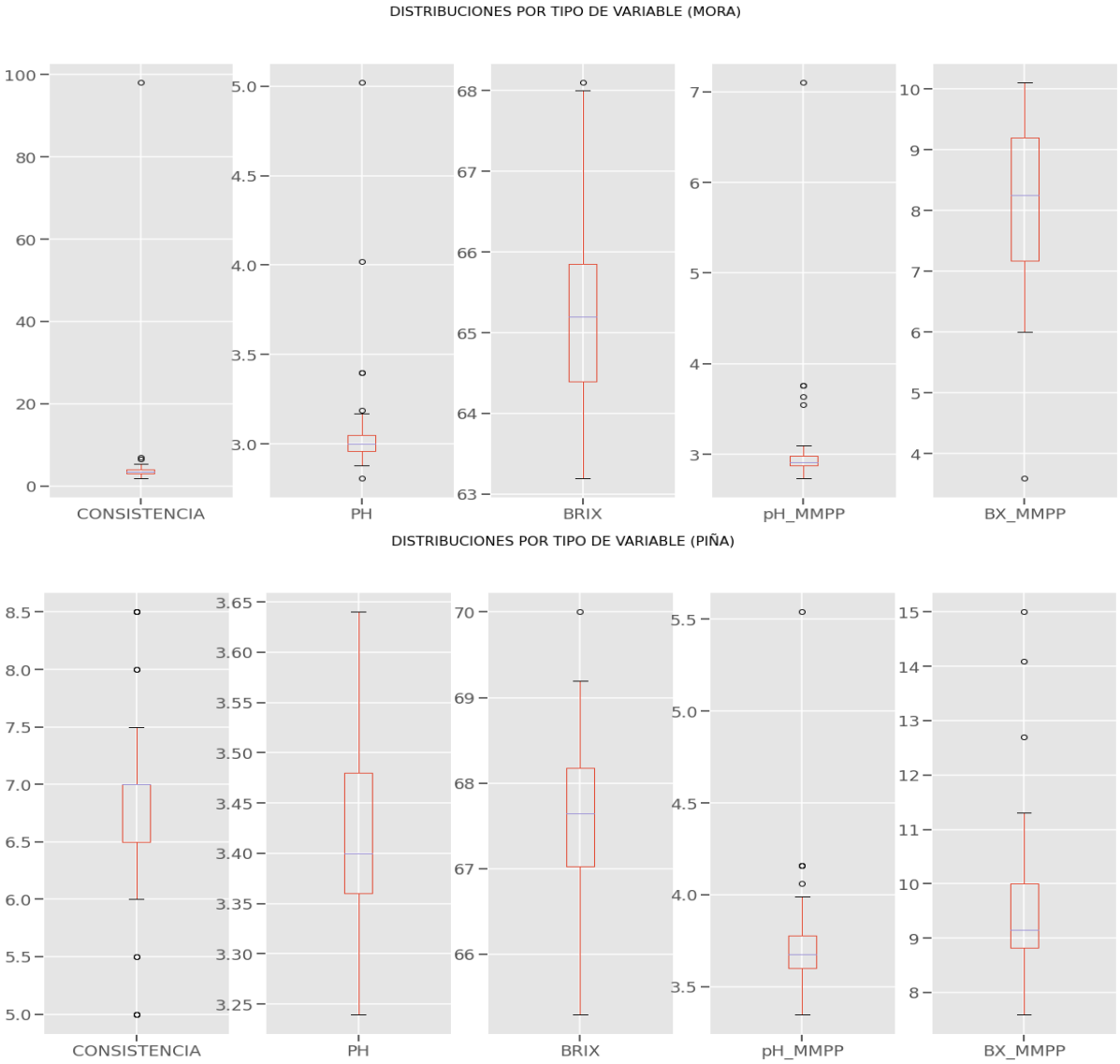
ANEXO D: DISTRIBUCIONES INICIALES DE LOS DATOS







ANEXO D: DISTRIBUCIONES INICIALES DE LOS DATOS (CONT.)



ANEXO E: CARACTERIZACIONES INICIALES DE LOS DATOS

FRUTILLA

	CONSISTENCIA	BRIX	PH	pH_MMPP	BX_MMPP
count	124.000000	125.000000	125.00000	122.000000	122.000000
mean	7.786290	66.112000	3.49560	3.567705	7.520492
std	1.261006	6.058273	0.11101	0.080451	0.704993
min	5.500000	1.000000	2.78000	3.290000	5.900000
25%	7.000000	65.400000	3.44000	3.522500	7.100000
50%	7.500000	67.000000	3.50000	3.580000	7.600000
75%	8.125000	67.700000	3.56000	3.610000	7.900000
max	11.500000	69.500000	3.68000	3.760000	9.700000

FRUTIMORA

	CONSISTENCIA	BRIX	PH	pH_FRUTIMORA	BX_FRUTIMORA
count	80.000000	80.000000	80.000000	78.000000	78.000000
mean	5.950000	65.916250	3.164625	3.211795	7.776795
std	0.919549	1.136889	0.071758	0.093943	0.682793
min	3.500000	64.000000	3.020000	2.840000	6.470000
25%	5.375000	65.100000	3.117500	3.170000	7.200000
50%	6.000000	65.600000	3.155000	3.210000	7.665000
75%	6.500000	66.800000	3.220000	3.220000	8.317500
max	7.500000	68.600000	3.350000	3.650000	9.070000

GUAYABA

	CONSISTENCIA	BRIX	PH	pH_MMPP	BX_MMPP	CONST
count	74.000000	74.000000	74.000000	72.000000	72.000000	65.000000
mean	6.608108	64.771622	3.442027	3.821806	15.573889	1.753846
std	2.671860	1.628557	0.168238	0.098729	4.320983	1.709631
min	0.500000	61.200000	3.230000	3.500000	8.000000	0.500000
25%	3.500000	63.550000	3.380000	3.787500	10.175000	0.500000
50%	7.750000	64.800000	3.420000	3.830000	18.100000	0.500000
75%	8.500000	65.700000	3.487500	3.880000	18.500000	3.500000
max	11.000000	68.500000	4.670000	4.000000	20.000000	5.000000

ANEXO E: CARACTERIZACIONES INICIALES DE LOS DATOS (CONT.)

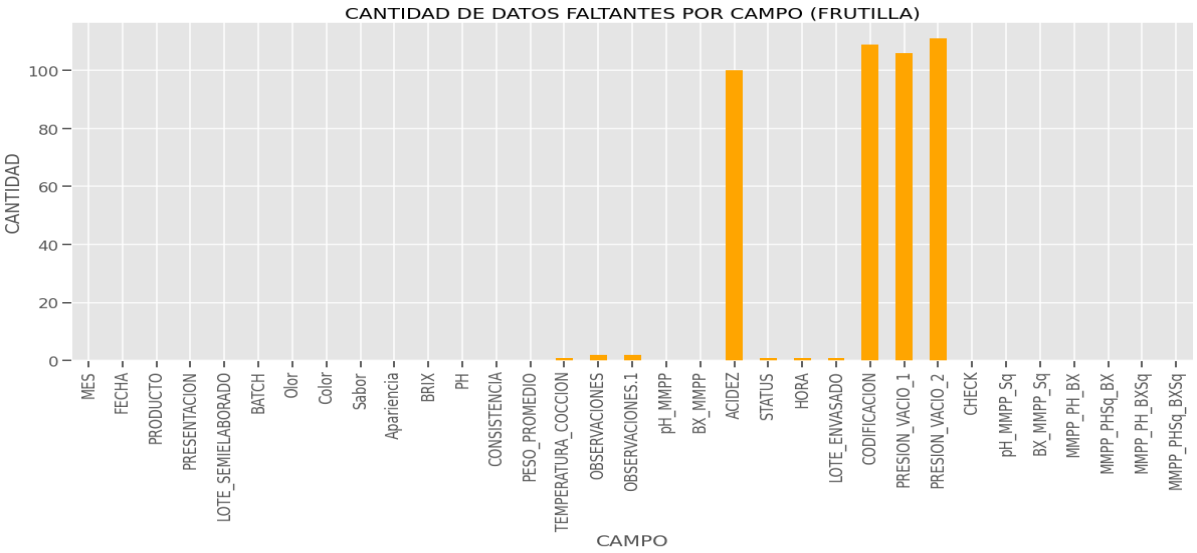
MORA

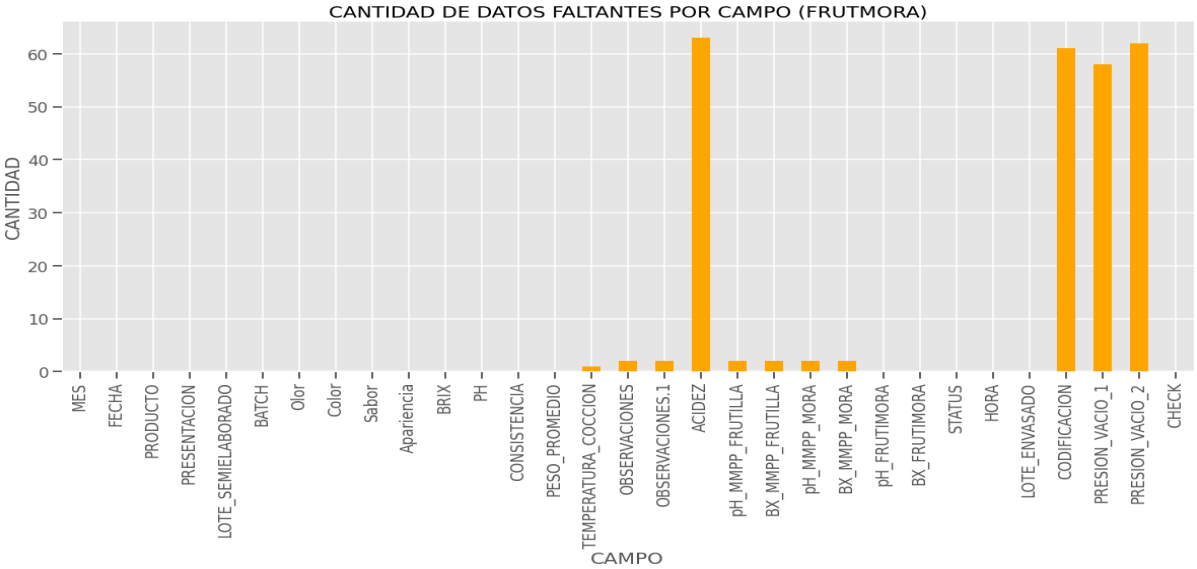
	CONSISTENCIA	BRIX	PH	pH_MMPP	BX_MMPP
count	104.000000	104.000000	104.000000	100.000000	100.000000
mean	4.500000	65.310577	3.038462	2.986100	8.206000
std	9.309667	1.027668	0.236716	0.447419	1.291919
min	2.000000	63.200000	2.810000	2.740000	3.600000
25%	3.000000	64.400000	2.960000	2.870000	7.100000
50%	3.500000	65.200000	3.000000	2.910000	8.250000
75%	4.000000	65.850000	3.050000	2.980000	9.225000
max	98.000000	68.100000	5.020000	7.100000	10.100000

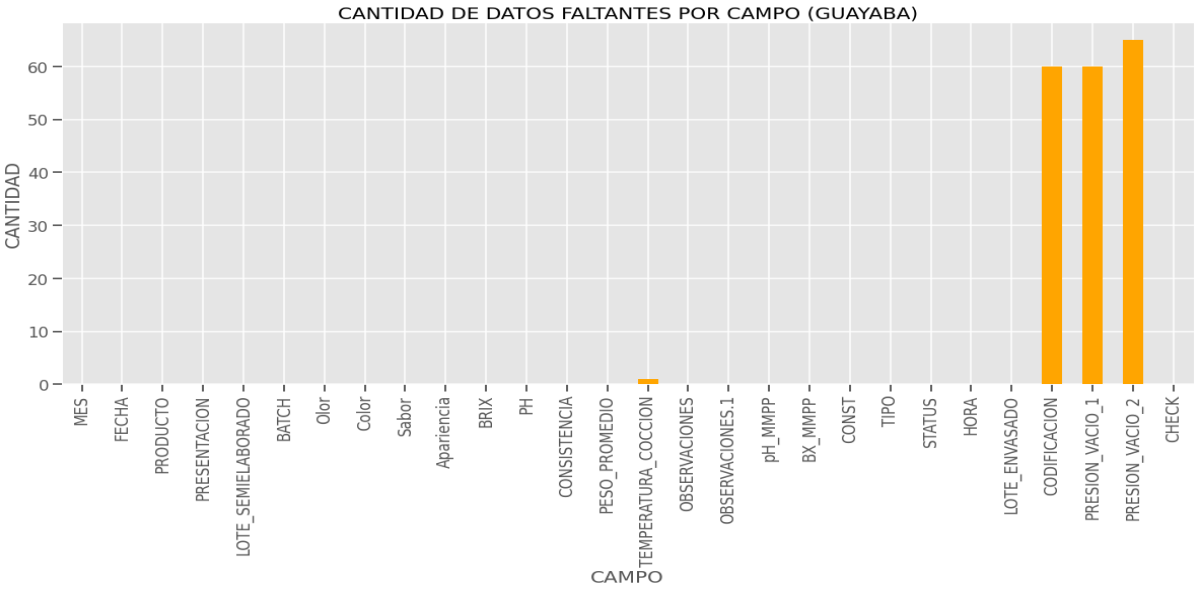
PIÑA

	CONSISTENCIA	BRIX	PH	pH_MMPP	BX_MMPP
count	70.000000	70.000000	70.000000	66.000000	66.000000
mean	6.792857	67.591429	3.421571	3.723636	9.459091
std	0.754181	0.869073	0.092871	0.277969	1.344960
min	5.000000	65.300000	3.240000	3.350000	7.600000
25%	6.500000	67.025000	3.360000	3.600000	8.800000
50%	7.000000	67.650000	3.400000	3.675000	9.150000
75%	7.000000	68.175000	3.480000	3.780000	10.000000
max	8.500000	70.000000	3.640000	5.540000	15.000000

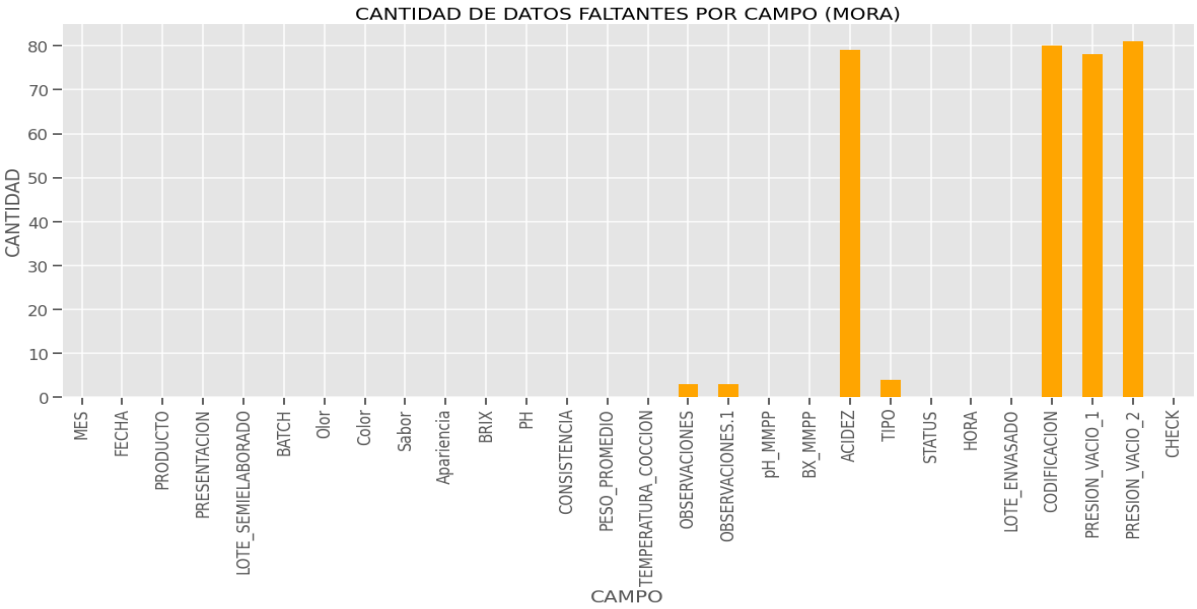
ANEXO F: CANTIDAD FINAL DE DATOS FALTANTES

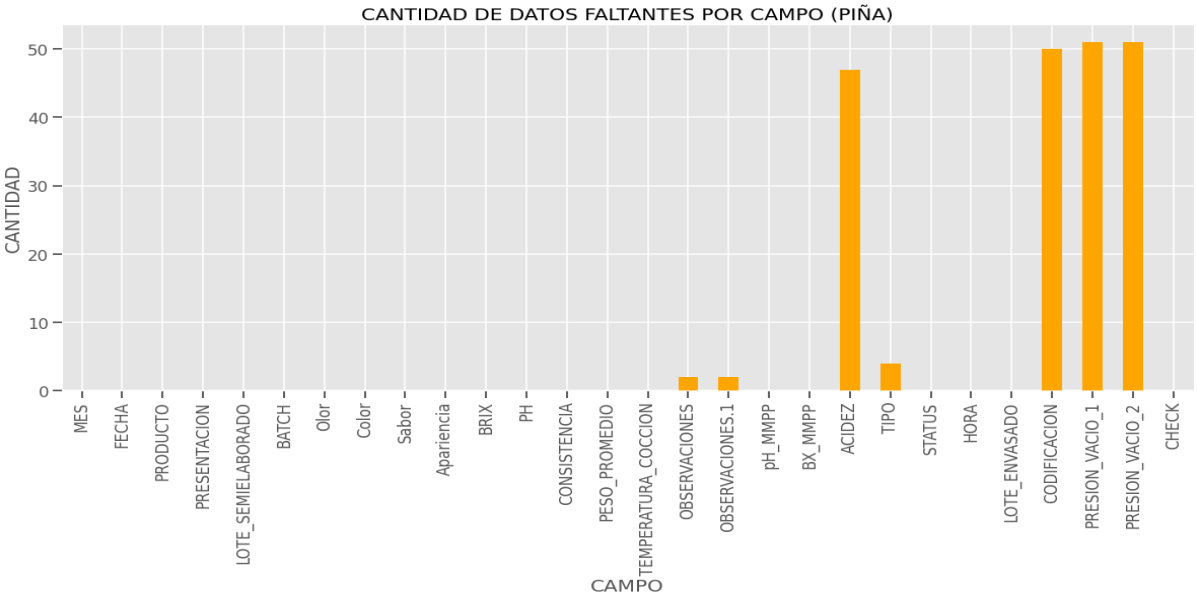




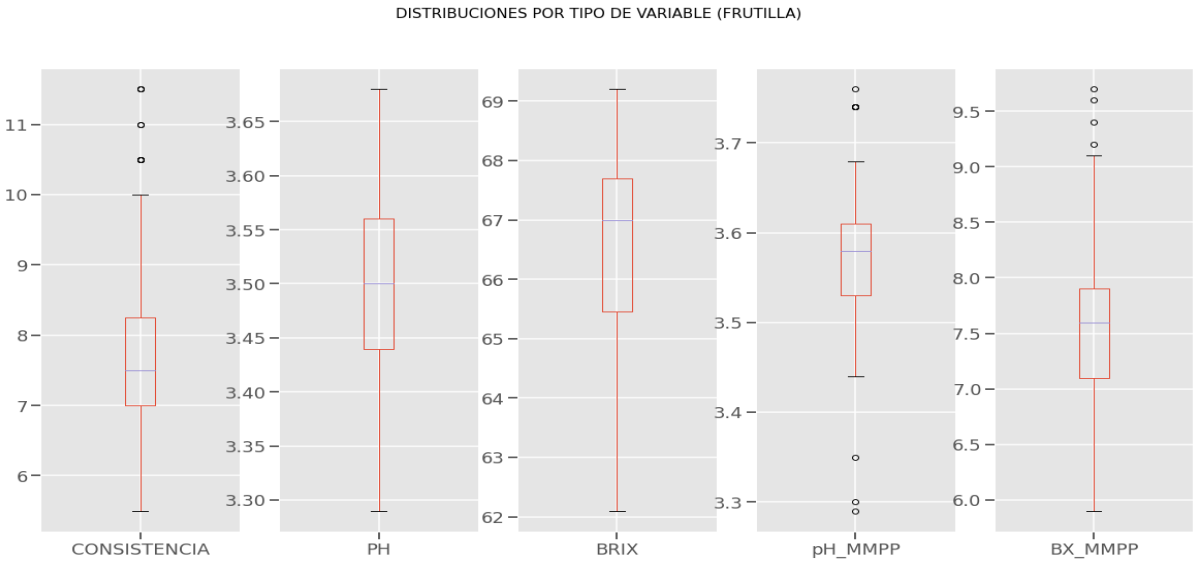


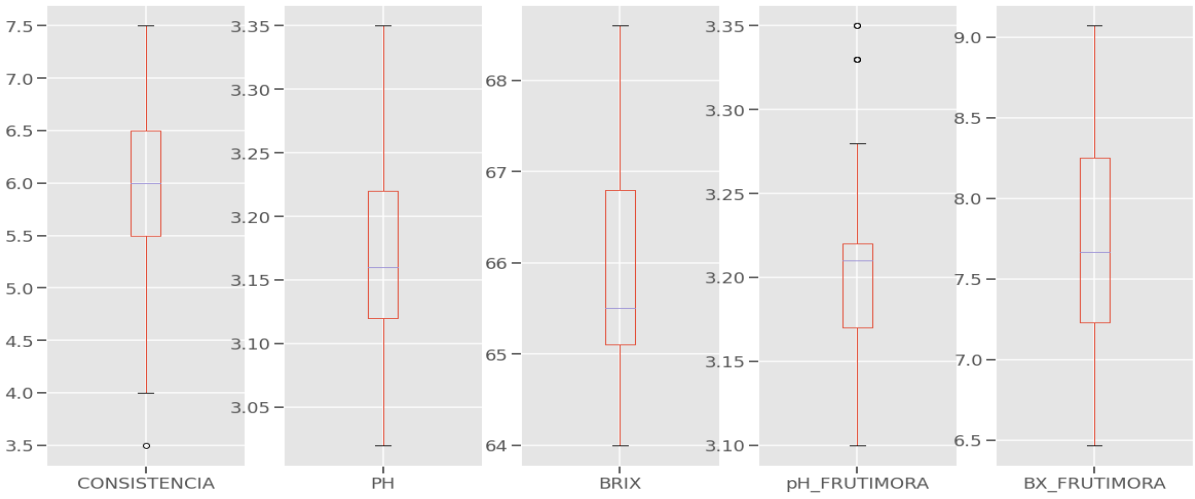
ANEXO F: CANTIDAD FINAL DE DATOS FALTANTES (CONT.)

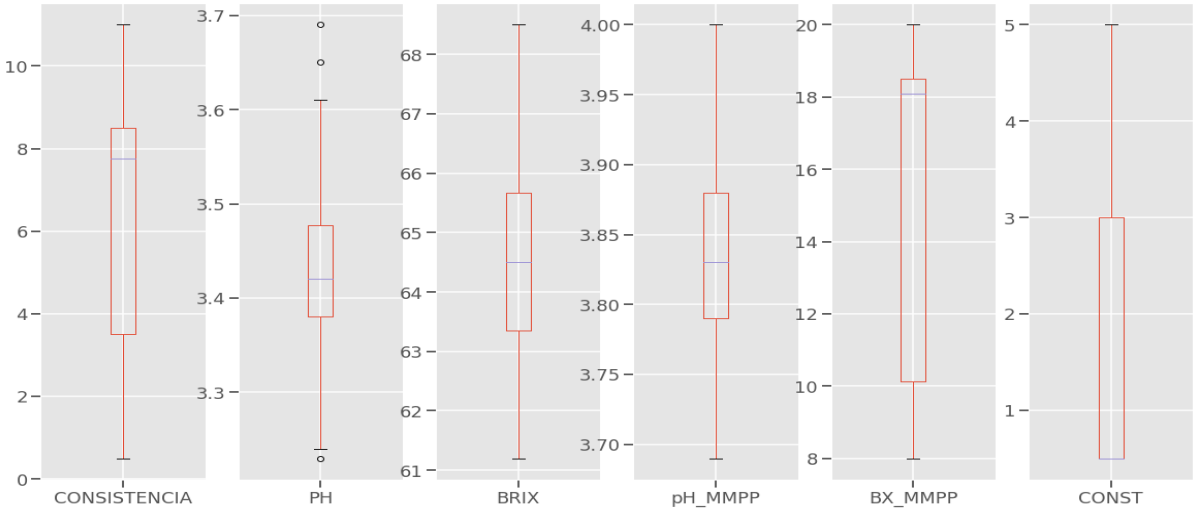




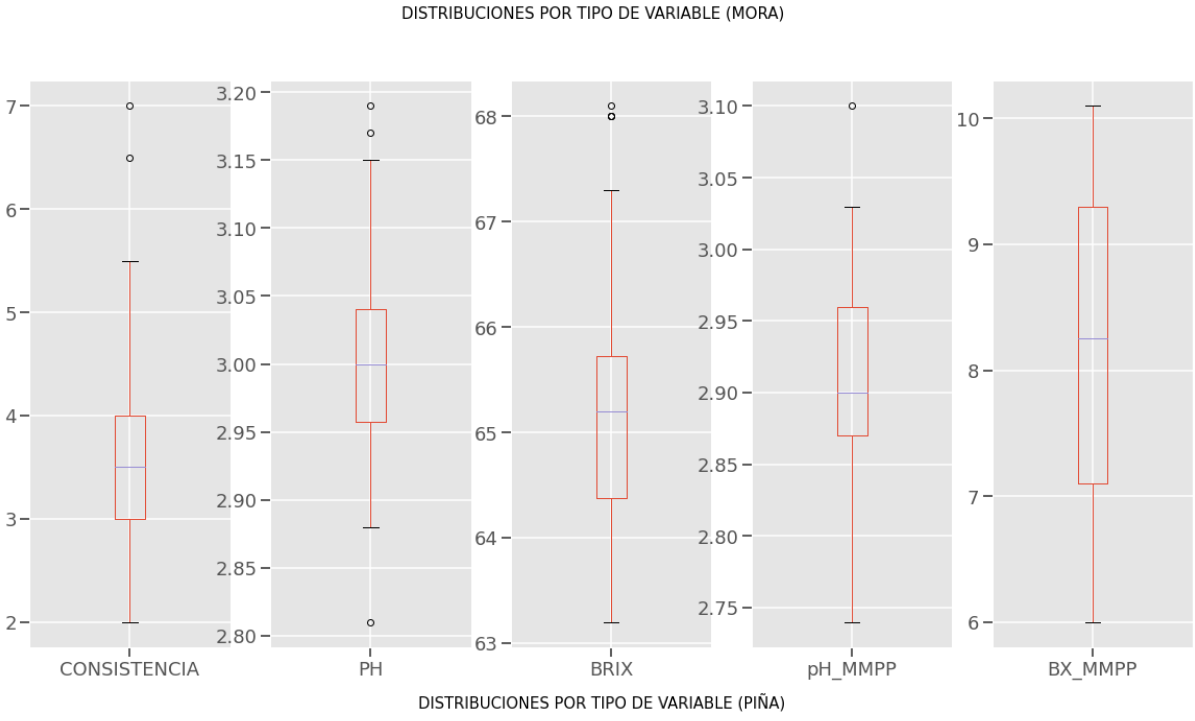
ANEXO G: DISTRIBUCIONES FINALES DE LOS DATOS

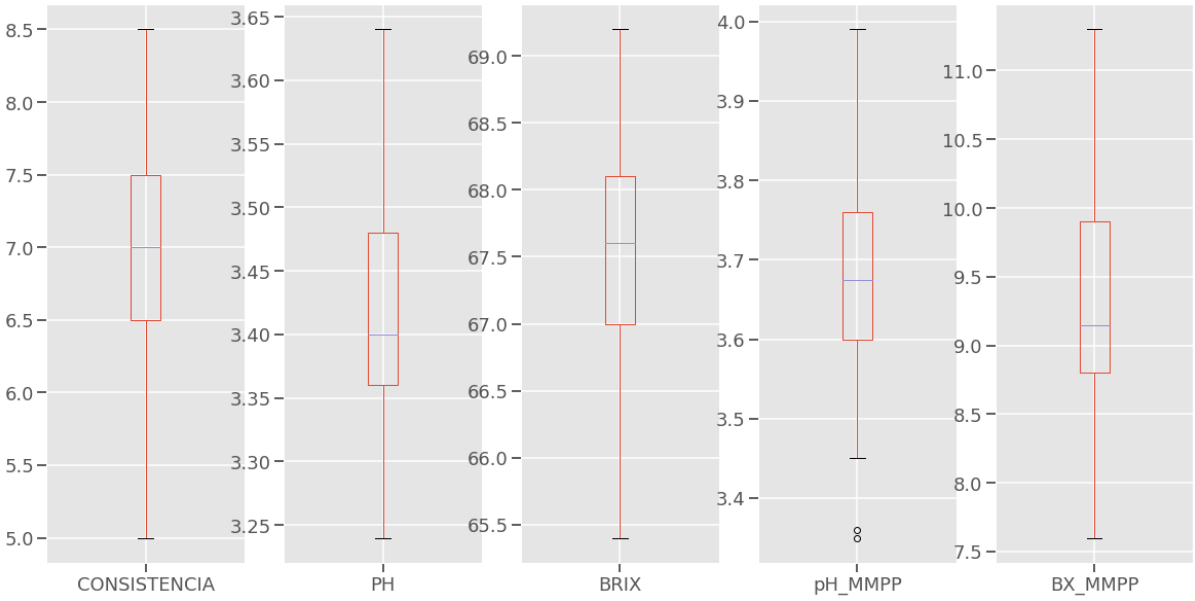






ANEXO G: DISTRIBUCIONES FINALES DE LOS DATOS (CONT.)





ANEXO H: CARACTERIZACIONES FINALES DE LOS DATOS

FRUTILLA

	pH_MMPP	BX_MMPP	BRIX	PH	CONSISTENCIA
count	123.000000	123.000000	123.000000	123.000000	123.000000
mean	3.569512	7.513821	66.613821	3.501301	7.792683
std	0.077606	0.698632	1.485191	0.091068	1.260898
min	3.290000	5.900000	62.100000	3.290000	5.500000
25%	3.530000	7.100000	65.450000	3.440000	7.000000
50%	3.580000	7.600000	67.000000	3.500000	7.500000
75%	3.610000	7.900000	67.700000	3.560000	8.250000
max	3.760000	9.700000	69.200000	3.680000	11.500000

FRUTIMORA

	pH_FRUTIMORA	BX_FRUTIMORA	CONSISTENCIA	PH	BRIX
count	77.000000	77.000000	77.000000	77.000000	77.00000
mean	3.205844	7.761039	5.935065	3.166883	65.88961
std	0.050767	0.669034	0.915332	0.071548	1.12444
min	3.100000	6.470000	3.500000	3.020000	64.00000
25%	3.170000	7.230000	5.500000	3.120000	65.10000
50%	3.210000	7.665000	6.000000	3.160000	65.50000
75%	3.220000	8.250000	6.500000	3.220000	66.80000
max	3.350000	9.070000	7.500000	3.350000	68.60000

GUAYABA

	pH_MMPP	BX_MMPP	CONST	BRIX	PH	CONSISTENCIA
count	70.000000	70.000000	70.000000	70.000000	70.000000	70.000000
mean	3.837143	15.548857	1.657143	64.724286	3.423857	6.542857
std	0.071892	4.384739	1.682267	1.659617	0.086582	2.726338
min	3.690000	8.000000	0.500000	61.200000	3.230000	0.500000
25%	3.790000	10.125000	0.500000	63.350000	3.380000	3.500000
50%	3.830000	18.100000	0.500000	64.500000	3.420000	7.750000
75%	3.880000	18.500000	3.000000	65.675000	3.477500	8.500000
max	4.000000	20.000000	5.000000	68.500000	3.690000	11.000000

ANEXO H: CARACTERIZACIONES FINALES DE LOS DATOS (CONT.)

MORA

	pH_MMPP	BX_MMPP	BRIX	PH	CONSISTENCIA
count	96.000000	96.000000	96.000000	96.000000	96.000000
mean	2.910104	8.252083	65.280208	3.000937	3.614583
std	0.066102	1.209043	1.031363	0.070189	0.993344
min	2.740000	6.000000	63.200000	2.810000	2.000000
25%	2.870000	7.100000	64.375000	2.957500	3.000000
50%	2.900000	8.250000	65.200000	3.000000	3.500000
75%	2.960000	9.300000	65.725000	3.040000	4.000000
max	3.100000	10.100000	68.100000	3.190000	7.000000

PIÑA

	pH_MMPP	BX_MMPP	BRIX	PH	CONSISTENCIA
count	61.000000	61.000000	61.000000	61.000000	61.000000
mean	3.670656	9.257377	67.554098	3.422787	6.811475
std	0.114766	0.912133	0.800744	0.091198	0.759188
min	3.350000	7.600000	65.400000	3.240000	5.000000
25%	3.600000	8.800000	67.000000	3.360000	6.500000
50%	3.675000	9.150000	67.600000	3.400000	7.000000
75%	3.760000	9.900000	68.100000	3.480000	7.500000
max	3.990000	11.300000	69.200000	3.640000	8.500000

ANEXO I: RESUMEN DEL ANÁLISIS EXPLORATORIO DE DATOS

	Registros Originales	Duplicados	Reprocesos	Atípicos	Registros Finales
Frutilla	166	31	10	2	123
Frutimora	95	15	0	3	77
Mora	111	6	1	8	96
Piña	104	26	8	9	61
Guayaba	91	13	4	4	70

ANEXO J: CORRELACIONES DE LOS DATOS

FRUTILLA

	CONSISTENCIA	BRIX	PH	pH_MMPP	BX_MMPP
CONSISTENCIA	1.000000	-0.637283	-0.353476	-0.131716	-0.166535
BRIX	-0.637283	1.000000	0.418328	0.255718	0.081813
PH	-0.353476	0.418328	1.000000	0.502279	0.095309
pH_MMPP	-0.131716	0.255718	0.502279	1.000000	0.226291
BX_MMPP	-0.166535	0.081813	0.095309	0.226291	1.000000

FRUTIMORA

	CONSISTENCIA	BRIX	PH	pH_FRUTIMORA	BX_FRUTIMORA
CONSISTENCIA	1.000000	-0.313876	0.035043	0.032343	0.106415
BRIX	-0.313876	1.000000	0.220550	-0.082594	-0.130605
PH	0.035043	0.220550	1.000000	0.078981	0.200029
pH_FRUTIMORA	0.032343	-0.082594	0.078981	1.000000	0.389118
BX_FRUTIMORA	0.106415	-0.130605	0.200029	0.389118	1.000000

GUAYABA

	CONSISTENCIA	BRIX	PH	pH_MMPP	BX_MMPP	CONST
CONSISTENCIA	1.000000	0.441146	-0.476531	0.191772	0.909605	-0.800160
BRIX	0.441146	1.000000	-0.250589	0.113069	0.623512	-0.558377
PH	-0.476531	-0.250589	1.000000	0.134043	-0.500129	0.434577
pH_MMPP	0.191772	0.113069	0.134043	1.000000	0.270682	-0.295215
BX_MMPP	0.909605	0.623512	-0.500129	0.270682	1.000000	-0.903810
CONST	-0.800160	-0.558377	0.434577	-0.295215	-0.903810	1.000000

MORA

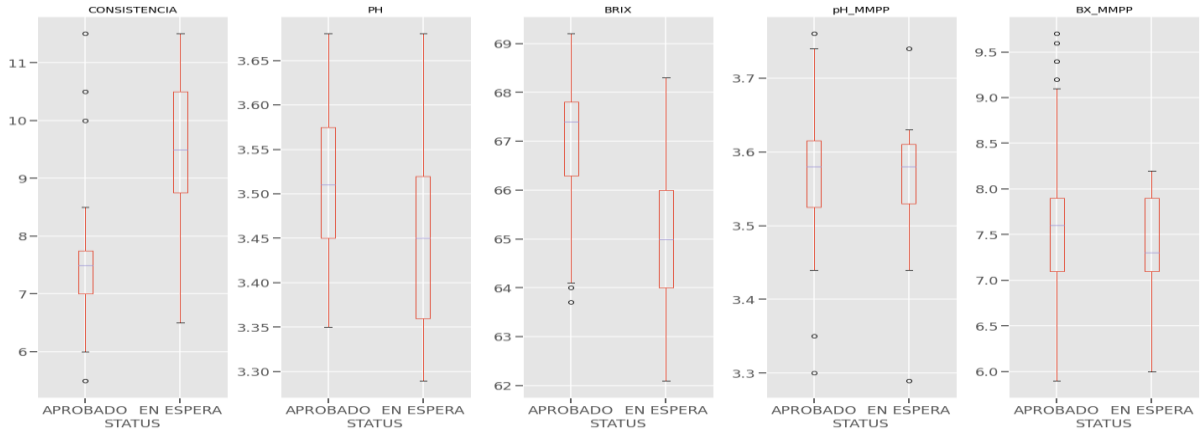
	CONSISTENCIA	BRIX	PH	pH_MMPP	BX_MMPP
CONSISTENCIA	1.000000	-0.112839	0.444578	-0.050682	0.047567
BRIX	-0.112839	1.000000	0.002004	-0.096779	-0.246038
PH	0.444578	0.002004	1.000000	0.205985	0.177046
pH_MMPP	-0.050682	-0.096779	0.205985	1.000000	0.352784
BX_MMPP	0.047567	-0.246038	0.177046	0.352784	1.000000

PIÑA

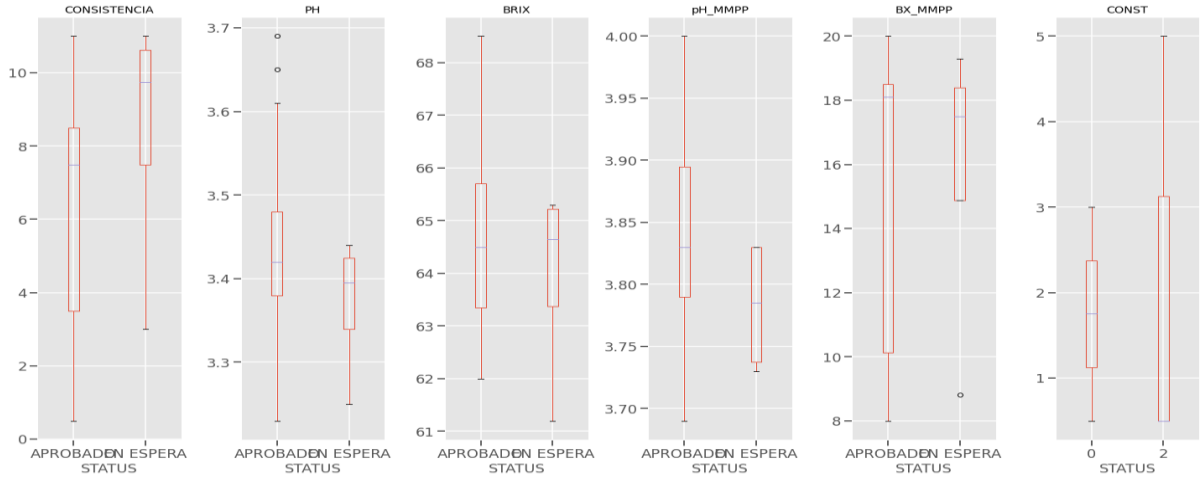
	CONSISTENCIA	BRIX	PH	pH_MMPP	BX_MMPP
CONSISTENCIA	1.000000	-0.019955	0.039009	0.428492	-0.171248
BRIX	-0.019955	1.000000	0.256483	-0.087536	-0.127201
PH	0.039009	0.256483	1.000000	0.353731	-0.402167
pH_MMPP	0.428492	-0.087536	0.353731	1.000000	-0.387252
BX_MMPP	-0.171248	-0.127201	-0.402167	-0.387252	1.000000

ANEXO K: ANÁLISIS CUALITATIVO DE LOS DATOS

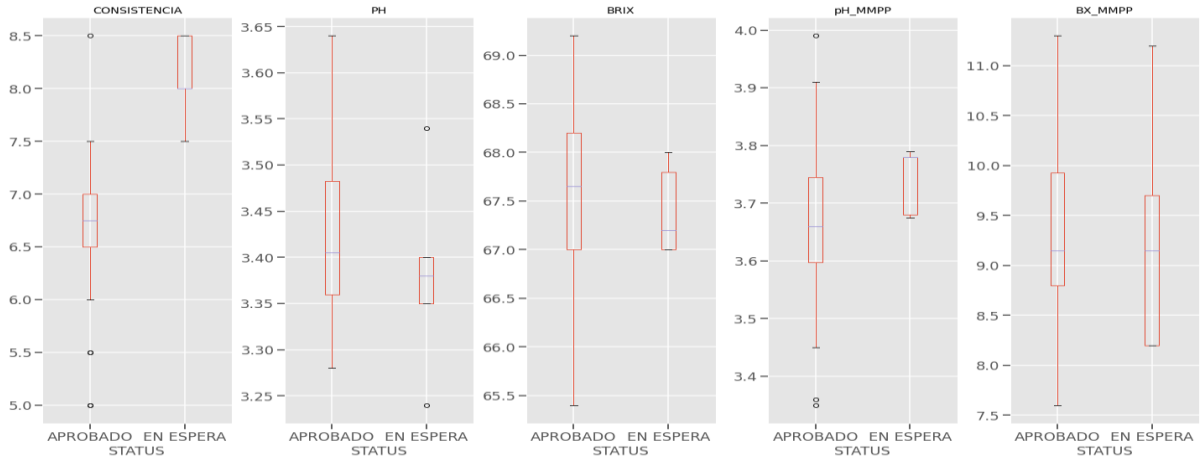
FRUTILLA



GUAYABA



PIÑA



ANEXO L: PRUEBAS SW DE NORMALIDAD

FRUTILLA

Shapiro-Wilk W test for normal data

Variable	Obs	W	V	z	Prob>z
ph	123	0.99329	0.659	-0.935	0.82501
brix	123	0.94897	5.014	3.617	0.00015
consistencia	123	0.90615	9.222	4.984	0.00000

FRUTIMORA

Shapiro-Wilk W test for normal data

Variable	Obs	W	V	z	Prob>z
brix	75	0.94337	3.687	2.849	0.00219
ph	75	0.98832	0.761	-0.597	0.72490
consistencia	75	0.97948	1.336	0.632	0.26371

GUAYABA

```
. swilk brix ph consistencia
```

Shapiro-Wilk W test for normal data

Variable	Obs	W	V	z	Prob>z
brix	70	0.97610	1.471	0.839	0.20064
ph	70	0.96756	1.997	1.504	0.06634
consistencia	70	0.89123	6.695	4.135	0.00002

MORA

Shapiro-Wilk W test for normal data

Variable	Obs	W	V	z	Prob>z
brix	96	0.94268	4.574	3.365	0.00038
ph	96	0.98512	1.188	0.381	0.35171
consistencia	96	0.96407	2.868	2.332	0.00986

PIÑA

Shapiro-Wilk W test for normal data

Variable	Obs	W	V	z	Prob>z
brix	61	0.98069	1.064	0.133	0.44696
ph	61	0.96505	1.925	1.413	0.07881
consistencia	61	0.99462	0.296	-2.626	0.99568

ANEXO M: TRANSFORMACIONES DE VARIABLES NO NORMALES

FRUTILLA

```
. ladder brix, gen (brix_t)
```

Transformation	formula	chi2(2)	P(chi2)
cubic	brix^3	7.09	0.029
square	brix^2	7.41	0.025
identity	brix	7.85	0.020
square root	sqrt(brix)	8.11	0.017
log	log(brix)	8.41	0.015
1/(square root)	1/sqrt(brix)	8.74	0.013
inverse	1/brix	9.11	0.011
1/square	1/(brix^2)	9.93	0.007
1/cubic	1/(brix^3)	10.88	0.004

```
(brix_t = brix^3   generated)

. ladder consistencia, gen (consistencia_t)
```

Transformation	formula	chi2(2)	P(chi2)
cubic	consis~a^3	40.80	0.000
square	consis~a^2	31.22	0.000
identity	consis~a	21.31	0.000
square root	sqrt(consis~a)	16.52	0.000
log	log(consis~a)	12.05	0.002
1/(square root)	1/sqrt(consis~a)	8.07	0.018
inverse	1/consis~a	4.68	0.096
1/square	1/(consis~a^2)	0.42	0.812
1/cubic	1/(consis~a^3)	6.24	0.044

```
(consistencia_t = 1/(consis~a^2)   generated)
```

FRUTIMORA

```
. ladder brix, gen(brix_t)
```

Transformation	formula	chi2(2)	P(chi2)
cubic	brix^3	5.16	0.076
square	brix^2	5.01	0.082
identity	brix	4.88	0.087
square root	sqrt(brix)	4.82	0.090
log	log(brix)	4.76	0.092
1/(square root)	1/sqrt(brix)	4.72	0.095
inverse	1/brix	4.67	0.097
1/square	1/(brix^2)	4.59	0.101
1/cubic	1/(brix^3)	4.53	0.104

```
(brix_t = 1/(brix^3)   generated)
```

ANEXO M: TRANSFORMACIONES DE VARIABLES NO NORMALES (CONT.)

GUAYABA

```
. ladder consistencia, gen(consistencia_t)

Transformation          formula                 chi2(2)       P(chi2)
-----------------------------------------------------------------------
cubic                   consis~a^3               3.27         0.195
square                  consis~a^2               8.46         0.015
identity                consis~a                15.30         0.000
square root             sqrt(consis~a)           7.08         0.029
log                     log(consis~a)           20.90         0.000
1/(square root)         1/sqrt(consis~a)        54.44         0.000
inverse                 1/consis~a                  .         0.000
1/square                1/(consis~a^2)              .         0.000
1/cubic                 1/(consis~a^3)              .         0.000

(consistencia_t = consis~a^3   generated)
```

MORA

```
. ladder brix, gen(brix_t)

Transformation          formula                 chi2(2)       P(chi2)
-----------------------------------------------------------------------
cubic                   brix^3                  10.41         0.005
square                  brix^2                   9.56         0.008
identity                brix                     8.75         0.013
square root             sqrt(brix)               8.35         0.015
log                     log(brix)                7.97         0.019
1/(square root)         1/sqrt(brix)             7.59         0.022
inverse                 1/brix                   7.22         0.027
1/square                1/(brix^2)               6.51         0.039
1/cubic                 1/(brix^3)               5.85         0.054

(brix_t = 1/(brix^3)   generated)

. ladder consistencia, gen(consistencia_t)

Transformation          formula                 chi2(2)       P(chi2)
-----------------------------------------------------------------------
cubic                   consis~a^3              53.95         0.000
square                  consis~a^2              30.12         0.000
identity                consis~a                 8.50         0.014
square root             sqrt(consis~a)           1.94         0.378
log                     log(consis~a)            0.47         0.792
1/(square root)         1/sqrt(consis~a)         2.28         0.319
inverse                 1/consis~a               6.13         0.047
1/square                1/(consis~a^2)          18.45         0.000
1/cubic                 1/(consis~a^3)          32.66         0.000

(consistencia_t = log(consis~a)   generated)
```

ANEXO N: PRUEBAS SW DE NORMALIDAD (VARIABLES TRANSFORMADAS)

FRUTILLA

Shapiro-Wilk W test for normal data

Variable	Obs	W	V	z	Prob>z
ph	123	0.99329	0.659	-0.935	0.82501
brix_t	123	0.95595	4.328	3.287	0.00051
consistenc~t	123	0.98804	1.175	0.362	0.35852

FRUTIMORA

Shapiro-Wilk W test for normal data

Variable	Obs	W	V	z	Prob>z
brix_t	75	0.95964	2.628	2.109	0.01746
ph	75	0.98832	0.761	-0.597	0.72490
consistencia	75	0.97948	1.336	0.632	0.26371

GUAYABA

Shapiro-Wilk W test for normal data

Variable	Obs	W	V	z	Prob>z
brix	70	0.97610	1.471	0.839	0.20064
ph	70	0.96756	1.997	1.504	0.06634
consistenc~t	70	0.91362	5.317	3.634	0.00014

MORA

Shapiro-Wilk W test for normal data

Variable	Obs	W	V	z	Prob>z
brix_t	96	0.95462	3.622	2.848	0.00220
ph	96	0.98512	1.188	0.381	0.35171
consistenc~t	96	0.98998	0.800	-0.495	0.68977

ANEXO O: NUEVAS VARIABLES

Variables Independientes	Variables Dependientes	Nuevas Variables	
Sólidos Solubles (BX_MMPP)	Sólidos Solubles (BRIX)	pH_MMPP^2	pH_MMPP^2* BX_MMPP
pH (pH_MMPP)	pH (PH)	BX_MMPP^2	pH_MMPP* BX_MMPP^2
	Consistencia (CONSISTENCIA)	pH_MMPP* BX_MMPP	pH_MMPP^2* BX_MMPP^2

ANEXO P: RESULTADOS MMR

FRUTILLA

Equation	Obs	Parms	RMSE	"R-sq"	F	P
ph	123	3	.0793817	0.2526	20.28246	0.0000
consistenc~t	123	3	.0047252	0.0363	2.259059	0.1089

	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
ph						
ph_mmpp	.594542	.0950733	6.25	0.000	.4063035	.7827806
bx_mmpp	-.0025214	.010561	-0.24	0.812	-.0234315	.0183887
_cons	1.398021	.3306495	4.23	0.000	.743358	2.052684
consistencia_t						
ph_mmpp	.00587	.0056593	1.04	0.302	-.0053349	.017075
bx_mmpp	.0009885	.0006286	1.57	0.118	-.0002561	.0022332
_cons	-.0108018	.019682	-0.55	0.584	-.0497707	.0281671

FRUTIMORA

. **mvreg**

Equation	Obs	Parms	RMSE	"R-sq"	F	P
ph	75	3	.0700802	0.0837	3.290308	0.0429
consistencia	75	3	.8958632	0.0751	2.924913	0.0601

	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
ph						
bx_mmpp_mora	-.2450896	.1287573	-1.90	0.061	-.5017627	.0115834
bx_mmpp_mora_sq	.0155639	.0077559	2.01	0.049	.0001027	.031025
_cons	4.114676	.5281141	7.79	0.000	3.0619	5.167452
consistencia						
bx_mmpp_mora	3.491963	1.645956	2.12	0.037	.2108102	6.773116
bx_mmpp_mora_sq	-.2027807	.0991471	-2.05	0.044	-.4004268	-.0051345
_cons	-8.829774	6.751091	-1.31	0.195	-22.28783	4.628279

GUAYABA

. **mvreg**

Equation	Obs	Parms	RMSE	"R-sq"	F	P
brix	70	4	1.323007	0.3921	14.19257	0.0000
ph	70	4	.0725448	0.3285	10.7623	0.0000

	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
brix						
ph_mmpp	-1.372798	2.318858	-0.59	0.556	-6.002547	3.256951
bx_mmpp	.2461111	.0848862	2.90	0.005	.0766303	.4155918
const	.0115932	.2229274	0.05	0.959	-.4334957	.4566822
_cons	66.14595	9.084754	7.28	0.000	48.00765	84.28425
ph						
ph_mmpp	.3491235	.1271505	2.75	0.008	.0952594	.6029876
bx_mmpp	-.0116966	.0046546	-2.51	0.014	-.0209898	-.0024034
const	-.0007828	.0122238	-0.06	0.949	-.0251884	.0236229
_cons	2.267386	.4981464	4.55	0.000	1.272805	3.261968

ANEXO P: RESULTADOS MMR (CONT.)

MORA

Equation	Obs	Parms	RMSE	"R-sq"	F	P
ph	96	9	.0660013	0.1902	2.554653	0.0150
consistenc~t	96	9	.2793452	0.0356	.4008702	0.9172

	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
ph						
ph_mmpp	-11.72967	10.34299	-1.13	0.260	-32.28748	8.82814
bx_mmpp	42.94141	13.284	3.23	0.002	16.53803	69.34479
ph_mmpp_sq	4.035105	3.595564	1.12	0.265	-3.111467	11.18168
bx_mmpp_sq	-5.484759	1.59665	-3.44	0.001	-8.658273	-2.311244
mmpp_ph_bx	-26.40578	8.137112	-3.25	0.002	-42.57917	-10.23239
mmpp_phsq_bx	3.999757	1.355497	2.95	0.004	1.305561	6.693953
mmpp_ph_bxsq	3.559118	1.029436	3.46	0.001	1.513001	5.605234
mmpp_phsq_bxsq	-.5750205	.16862	-3.41	0.001	-.9101709	-.23987
_cons	3.000938	.	.	.	.	.
consistencia t						
ph_mmpp	.7775405	43.77585	0.02	0.986	-86.23169	87.78677
bx_mmpp	22.42622	56.22342	0.40	0.691	-89.32392	134.1764
ph_mmpp_sq	-.1233724	15.21793	-0.01	0.994	-30.37065	30.12391
bx_mmpp_sq	-3.842185	6.757689	-0.57	0.571	-17.27382	9.589454
mmpp_ph_bx	-15.19579	34.43966	-0.44	0.660	-83.64834	53.25676
mmpp_phsq_bx	2.536518	5.737028	0.44	0.659	-8.866446	13.93948
mmpp_ph_bxsq	2.597373	4.357005	0.60	0.553	-6.062646	11.25739
mmpp_phsq_bxsq	-.436473	.7136702	-0.61	0.542	-1.85497	.9820238
_cons	1.248435	.	.	.	.	.

PIÑA

Equation	Obs	Parms	RMSE	"R-sq"	F	P
brix	61	6	.8106845	0.0604	.707521	0.6203
ph	61	6	.0807457	0.2814	4.307879	0.0022
consistencia	61	6	.6661914	0.2942	4.584096	0.0014

	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
brix						
ph_mmpp	21.95138	42.07777	0.52	0.604	-62.37435	106.2771
bx_mmpp	-.7805462	6.079478	-0.13	0.898	-12.96409	11.403
ph_mmpp_sq	-2.711907	4.954842	-0.55	0.586	-12.64163	7.217819
bx_mmpp_sq	.0959337	.1219612	0.79	0.435	-.1484821	.3403495
mmpp_ph_bx	-.3215248	1.289569	-0.25	0.804	-2.905879	2.26283
_cons	33.39139	95.36576	0.35	0.728	-157.7259	224.5086
ph						
ph_mmpp	-5.443998	4.191023	-1.30	0.199	-13.843	2.955
bx_mmpp	.5088999	.6055272	0.84	0.404	-.7046038	1.722404
ph_mmpp_sq	.8663902	.4935114	1.76	0.085	-.1226289	1.855409
bx_mmpp_sq	-.0135152	.0121476	-1.11	0.271	-.0378595	.0108291
mmpp_ph_bx	-.0798267	.1284435	-0.62	0.537	-.3372332	.1775798
_cons	10.88871	9.498606	1.15	0.257	-8.146918	29.92435
consistencia						
ph_mmpp	65.75864	34.578	1.90	0.062	-3.537214	135.0545
bx_mmpp	-4.457362	4.995897	-0.89	0.376	-14.46936	5.554638
ph_mmpp_sq	-9.543188	4.071711	-2.34	0.023	-17.70308	-1.383296
bx_mmpp_sq	.0907807	.1002234	0.91	0.369	-.1100714	.2916328
mmpp_ph_bx	.7653663	1.059722	0.72	0.473	-1.358364	2.889096
_cons	-98.4278	78.36816	-1.26	0.214	-255.4811	58.6255

ANEXO Q: RESULTADOS MANOVA

FRUTILLA

Number of obs = 123

W = Wilks' lambda L = Lawley-Hotelling trace
P = Pillai's trace R = Roy's largest root

Source	Statistic		df	F(df1,	df2) =	F	Prob>F	
Model	W	0.7276	2	4.0	238.0	10.25	0.0000	e
	P	0.2784		4.0	240.0	9.70	0.0000	a
	L	0.3661		4.0	236.0	10.80	0.0000	a
	R	0.3420		2.0	120.0	20.52	0.0000	u
Residual			120					
ph_mmpp	W	0.7497	1	2.0	119.0	19.86	0.0000	e
	P	0.2503		2.0	119.0	19.86	0.0000	e
	L	0.3338		2.0	119.0	19.86	0.0000	e
	R	0.3338		2.0	119.0	19.86	0.0000	e
bx_mmpp	W	0.9751	1	2.0	119.0	1.52	0.2227	e
	P	0.0249		2.0	119.0	1.52	0.2227	e
	L	0.0256		2.0	119.0	1.52	0.2227	e
	R	0.0256		2.0	119.0	1.52	0.2227	e
Residual			120					
Total			122					

e = exact, a = approximate, u = upper bound on F

FRUTIMORA

Number of obs = 75

W = Wilks' lambda L = Lawley-Hotelling trace
P = Pillai's trace R = Roy's largest root

Source	Statistic		df	F(df1,	df2) =	F	Prob>F	
Model	W	0.8453	2	4.0	142.0	3.11	0.0172	e
	P	0.1604		4.0	144.0	3.14	0.0165	a
	L	0.1764		4.0	140.0	3.09	0.0180	a
	R	0.1220		2.0	72.0	4.39	0.0159	u
Residual			72					
bx_mmpp_~ra	W	0.8932	1	2.0	71.0	4.25	0.0181	e
	P	0.1068		2.0	71.0	4.25	0.0181	e
	L	0.1196		2.0	71.0	4.25	0.0181	e
	R	0.1196		2.0	71.0	4.25	0.0181	e
bx_mmpp_m~q	W	0.8921	1	2.0	71.0	4.29	0.0174	e
	P	0.1079		2.0	71.0	4.29	0.0174	e
	L	0.1209		2.0	71.0	4.29	0.0174	e
	R	0.1209		2.0	71.0	4.29	0.0174	e
Residual			72					
Total			74					

e = exact, a = approximate, u = upper bound on F

ANEXO Q: RESULTADOS MANOVA (CONT.)

GUAYABA

Number of obs = 70

W = Wilks' lambda L = Lawley-Hotelling trace
P = Pillai's trace R = Roy's largest root

Source	Statistic	df	F(df1, df2) = F	Prob>F
Model	W 0.4291	3	6.0 130.0	11.41 0.0000 e
	P 0.5935		6.0 132.0	9.28 0.0000 a
	L 1.2776		6.0 128.0	13.63 0.0000 a
	R 1.2350		3.0 66.0	27.17 0.0000 u
Residual		66		
ph_mmpp	W 0.8870	1	2.0 65.0	4.14 0.0203 e
	P 0.1130		2.0 65.0	4.14 0.0203 e
	L 0.1274		2.0 65.0	4.14 0.0203 e
	R 0.1274		2.0 65.0	4.14 0.0203 e
bx_mmpp	W 0.7977	1	2.0 65.0	8.24 0.0006 e
	P 0.2023		2.0 65.0	8.24 0.0006 e
	L 0.2535		2.0 65.0	8.24 0.0006 e
	R 0.2535		2.0 65.0	8.24 0.0006 e
const	W 0.9999	1	2.0 65.0	0.00 0.9962 e
	P 0.0001		2.0 65.0	0.00 0.9962 e
	L 0.0001		2.0 65.0	0.00 0.9962 e
	R 0.0001		2.0 65.0	0.00 0.9962 e
Residual		66		
Total		69		

MORA

ph_mmpp	W 0.9818	1	2.0	86.0	0.80 0.4537 e
	P 0.0182		2.0	86.0	0.80 0.4537 e
	L 0.0186		2.0	86.0	0.80 0.4537 e
	R 0.0186		2.0	86.0	0.80 0.4537 e
bx_mmpp	W 0.8812	1	2.0	86.0	5.80 0.0043 e
	P 0.1188		2.0	86.0	5.80 0.0043 e
	L 0.1348		2.0	86.0	5.80 0.0043 e
	R 0.1348		2.0	86.0	5.80 0.0043 e
ph_mmpp_sq	W 0.9823	1	2.0	86.0	0.78 0.4637 e
	P 0.0177		2.0	86.0	0.78 0.4637 e
	L 0.0180		2.0	86.0	0.78 0.4637 e
	R 0.0180		2.0	86.0	0.78 0.4637 e
bx_mmpp_sq	W 0.8710	1	2.0	86.0	6.37 0.0026 e
	P 0.1290		2.0	86.0	6.37 0.0026 e
	L 0.1481		2.0	86.0	6.37 0.0026 e
	R 0.1481		2.0	86.0	6.37 0.0026 e
mmpp_ph_bx	W 0.8813	1	2.0	86.0	5.79 0.0044 e
	P 0.1187		2.0	86.0	5.79 0.0044 e
	L 0.1347		2.0	86.0	5.79 0.0044 e
	R 0.1347		2.0	86.0	5.79 0.0044 e
mmpp_phsq~x	W 0.9006	1	2.0	86.0	4.75 0.0111 e
	P 0.0994		2.0	86.0	4.75 0.0111 e
	L 0.1104		2.0	86.0	4.75 0.0111 e
	R 0.1104		2.0	86.0	4.75 0.0111 e
mmpp_ph_b~q	W 0.8700	1	2.0	86.0	6.42 0.0025 e
	P 0.1300		2.0	86.0	6.42 0.0025 e
	L 0.1494		2.0	86.0	6.42 0.0025 e
	R 0.1494		2.0	86.0	6.42 0.0025 e
mmpp_phsq~q	W 0.8736	1	2.0	86.0	6.22 0.0030 e
	P 0.1264		2.0	86.0	6.22 0.0030 e
	L 0.1448		2.0	86.0	6.22 0.0030 e
	R 0.1448		2.0	86.0	6.22 0.0030 e

ANEXO Q: RESULTADOS MANOVA (CONT.)

PIÑA

Number of obs = 61

W = Wilks' lambda L = Lawley-Hotelling trace
P = Pillai's trace R = Roy's largest root

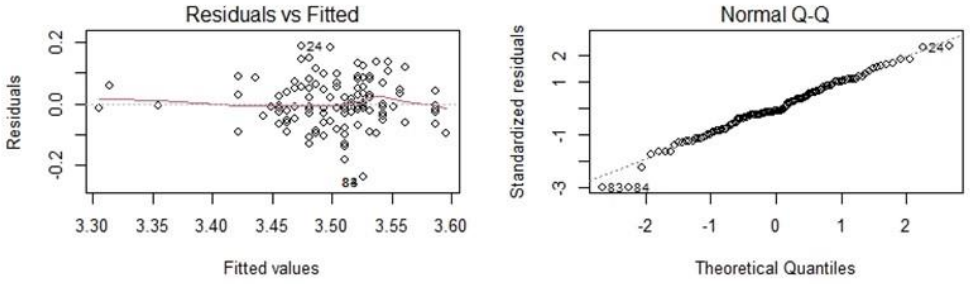
Source	Statistic		df	F(df1,	df2) =	F	Prob>F	
Model	W	0.4583	5	15.0	146.7	3.20	0.0001	a
	P	0.6585		15.0	165.0	3.09	0.0002	a
	L	0.9384		15.0	155.0	3.23	0.0001	a
	R	0.5550		5.0	55.0	6.10	0.0001	u
Residual			55					
ph_mmpp	W	0.9000	1	3.0	53.0	1.96	0.1306	e
	P	0.1000		3.0	53.0	1.96	0.1306	e
	L	0.1112		3.0	53.0	1.96	0.1306	e
	R	0.1112		3.0	53.0	1.96	0.1306	e
bx_mmpp	W	0.9713	1	3.0	53.0	0.52	0.6687	e
	P	0.0287		3.0	53.0	0.52	0.6687	e
	L	0.0296		3.0	53.0	0.52	0.6687	e
	R	0.0296		3.0	53.0	0.52	0.6687	e
ph_mmpp_sq	W	0.8506	1	3.0	53.0	3.10	0.0342	e
	P	0.1494		3.0	53.0	3.10	0.0342	e
	L	0.1757		3.0	53.0	3.10	0.0342	e
	R	0.1757		3.0	53.0	3.10	0.0342	e
bx_mmpp_sq	W	0.9411	1	3.0	53.0	1.11	0.3547	e
	P	0.0589		3.0	53.0	1.11	0.3547	e
	L	0.0626		3.0	53.0	1.11	0.3547	e
	R	0.0626		3.0	53.0	1.11	0.3547	e
mmpp_ph_bx	W	0.9843	1	3.0	53.0	0.28	0.8378	e
	P	0.0157		3.0	53.0	0.28	0.8378	e
	L	0.0160		3.0	53.0	0.28	0.8378	e
	R	0.0160		3.0	53.0	0.28	0.8378	e
Residual			55					
Total			60					

e = exact, a = approximate, u = upper bound on F

ANEXO R: ANÁLISIS DE RESIDUALES

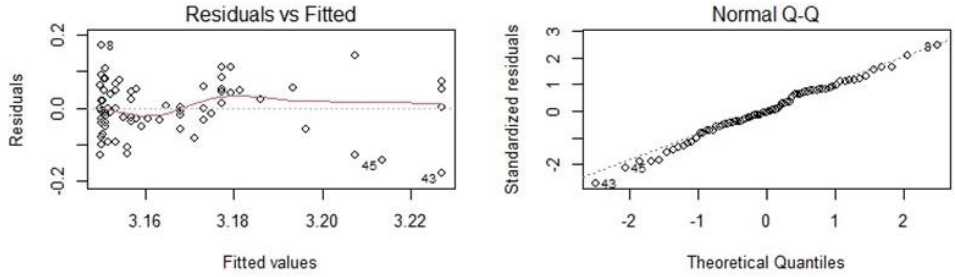
FRUTILLA

pH



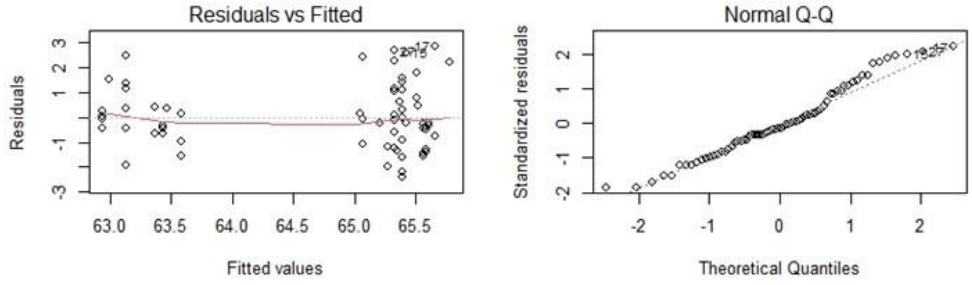
FRUTIMORA

pH

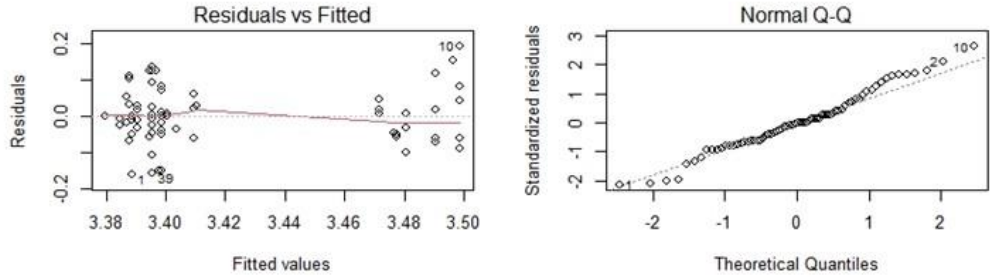


GUAYABA

Brix



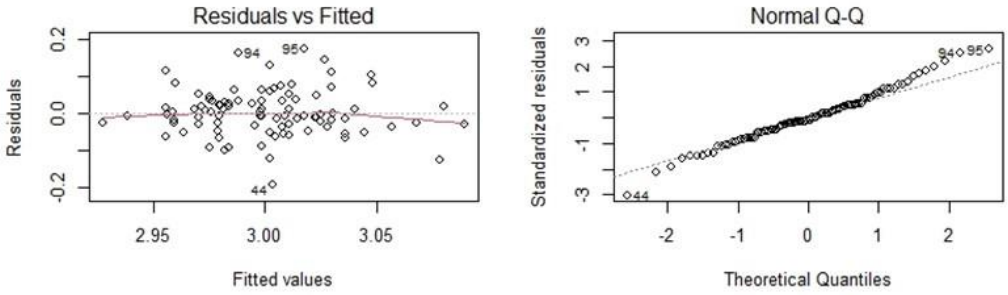
pH



ANEXO R: ANÁLISIS DE RESIDUALES (CONT.)

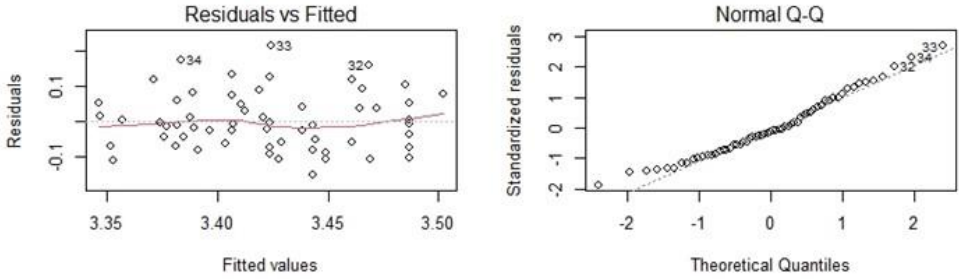
MORA

pH

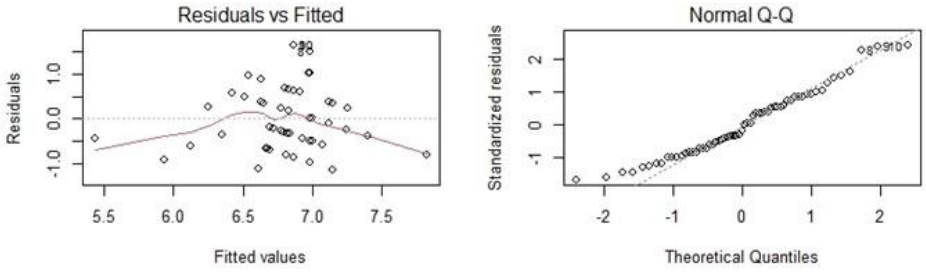


PIÑA

pH



Const



ANEXO S: RESULTADOS GLM

FRUTILLA

Coefficients: glm coefficients

	names	coefficients	std_error	z_value	p_value	standardized_coefficients
1	Intercept	49.132305	6.028282	8.150300	0.000000	66.613821
2	pH_MMPP	4.784511	1.733342	2.760280	0.006682	0.371307
3	BX_MMPP	0.053654	0.192545	0.278657	0.780988	0.037484

```
MSE:  2.043468
RMSE:  1.429499
MAE:  1.168735
RMSLE:  0.02129497
Mean Residual Deviance :  2.043468
R^2 :  0.06599614
Null Deviance :269.1065
Null D.o.F. :122
Residual Deviance :251.3465
Residual D.o.F. :120
AIC :444.9606
```

FRUTIMORA

Coefficients: glm coefficients

	names	coefficients	std_error	z_value	p_value	standardized_coefficients
1	Intercept	84.183587	22.335499	3.769049	0.000349	65.862667
2	pH_FRUTIMORA	-61.013252	48.238745	-1.264818	0.210318	-3.138724
3	BX_FRUTIMORA	-13.264257	55.506628	-0.238967	0.811861	-8.990854
4	pH_MMPP_FRUTILLA	25.403963	21.819323	1.164287	0.248437	2.913038
5	pH_MMPP_MORA	32.997286	26.432457	1.248362	0.216243	3.458136
6	BX_MMPP_FRUTILLA	4.843295	25.280188	0.191585	0.848647	2.964653
7	BX_MMPP_MORA	5.936970	30.856888	0.192403	0.848008	6.043553
8	BX_MMPP_MORA_FRUTILLA	0.166771	0.288787	0.577488	0.565545	1.647152

```
MSE:  1.137296
RMSE:  1.066441
MAE:  0.8557423
RMSLE:  0.01588283
Mean Residual Deviance :  1.137296
R^2 :  0.08943918
Null Deviance :93.67548
Null D.o.F. :74
Residual Deviance :85.29722
Residual D.o.F. :67
AIC :240.4898
```

ANEXO S: RESULTADOS GLM (CONT.)

GUAYABA

Coefficients: glm coefficients

	names	coefficients	std_error	z_value	p_value	standardized_coefficients
1	Intercept	-2.251161	0.505838	-4.450362	0.000033	6.542857
2	BX_MMPP	0.565573	0.031327	18.053583	0.000000	2.479891

MSE: 1.264731
RMSE: 1.124603
MAE: 0.7949662
RMSLE: 0.1776481
Mean Residual Deviance : 1.264731
R^2 : 0.8273813
Null Deviance :512.8714
Null D.o.F. :69
Residual Deviance :88.53119
Residual D.o.F. :68
AIC :221.0916

MORA

Coefficients: glm coefficients

	names	coefficients	std_error	z_value	p_value	standardized_coefficients
1	Intercept	839.125199	1359.242889	0.617347	0.538600	3.614583
2	pH_MMPP	-593.584983	984.344574	-0.603026	0.548043	-39.237102
3	BX_MMPP	-125.577311	169.199292	-0.742186	0.459951	-151.828374
4	pH_MMPP_Sq	106.100480	181.707661	0.583908	0.560776	40.929420
5	BX_MMPP_Sq	-1.481825	7.072714	-0.209513	0.834532	-29.171013
6	MMPP_PH_BX	92.380536	117.257152	0.787846	0.432904	343.472512
7	MMPP_PHSq_BX	-17.119242	22.489788	-0.761201	0.448572	-198.204704
8	MMPP_PH_BXSq	0.546872	2.455903	0.222676	0.824303	32.034838
9	MMPP_PHSq_BXSq	0.000000	NA	NA	NA	0.000000

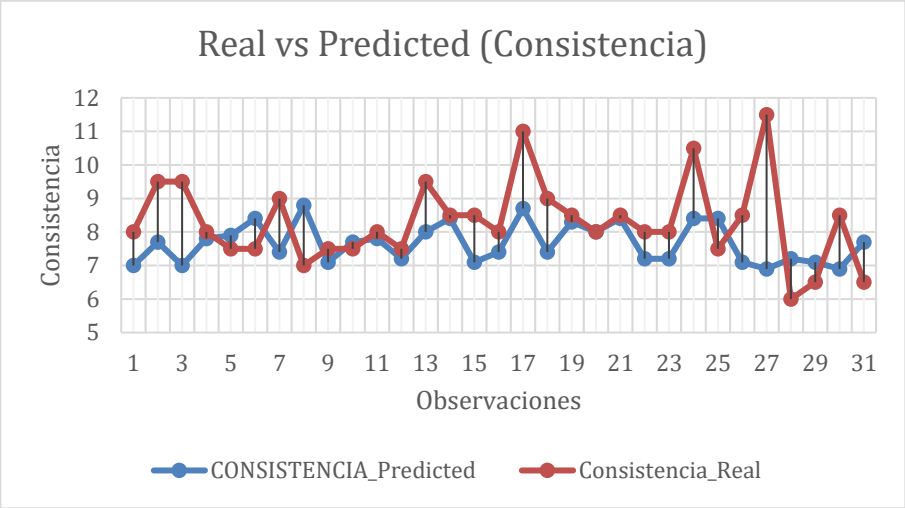
MSE: 0.9302543
RMSE: 0.9644969
MAE: 0.7704137
RMSLE: 0.2070001
Mean Residual Deviance : 0.9302543
R^2 : 0.04731378
Null Deviance :93.73958
Null D.o.F. :95
Residual Deviance :89.30441
Residual D.o.F. :88
AIC :283.4957

ANEXO T: ERRORES MEDIOS ABSOLUTOS

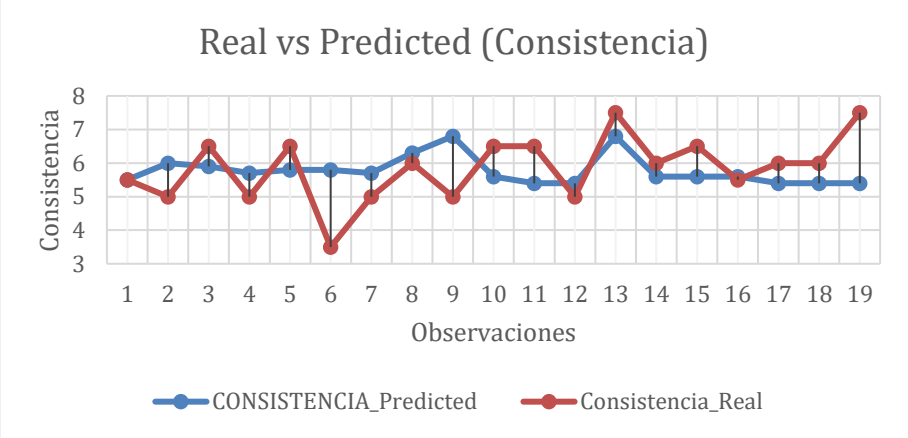
	MAE (Decision Tree)	MAE (KNN)
Frutilla	0.91	0.77
Frutimora	0.71	0.58
Mora	0.63	0.55
Guayaba	0.47	0.59
Piña	0.53	0.51

ANEXO U: GRÁFICOS DATOS REALES VERSUS PREDICHOS

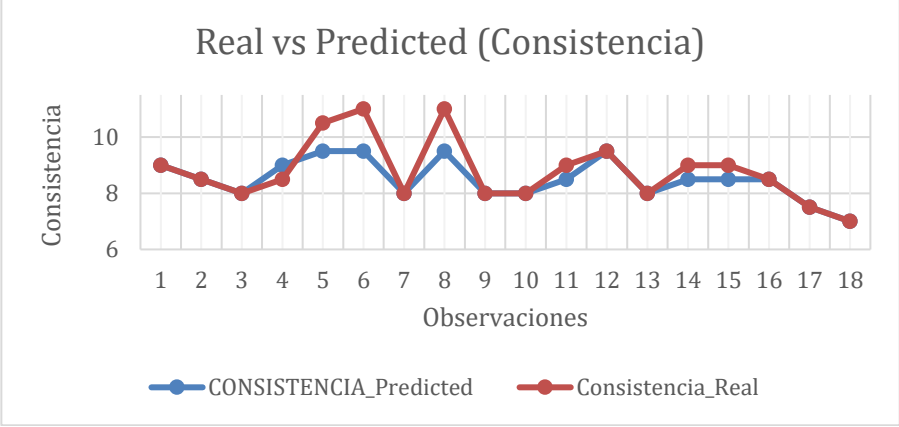
FRUTILLA



FRUTIMORA



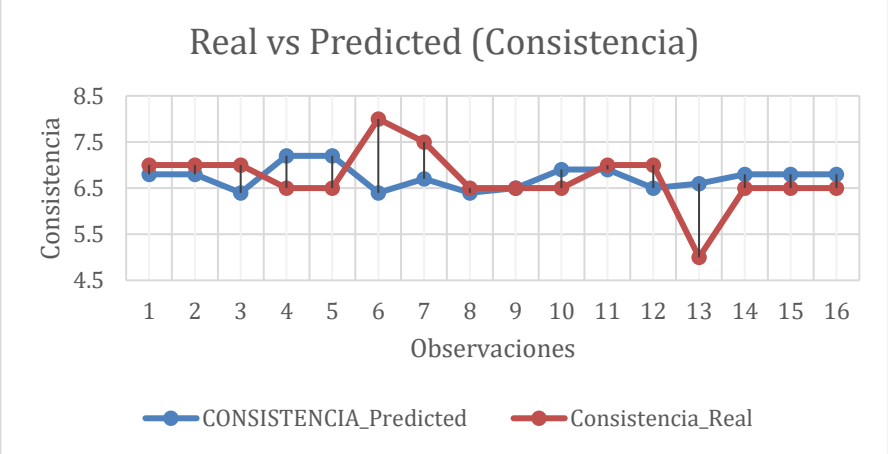
GUAYABA



ANEXO U: GRÁFICOS DE DATOS REALES VERSUS PREDICHOS (CONT.)

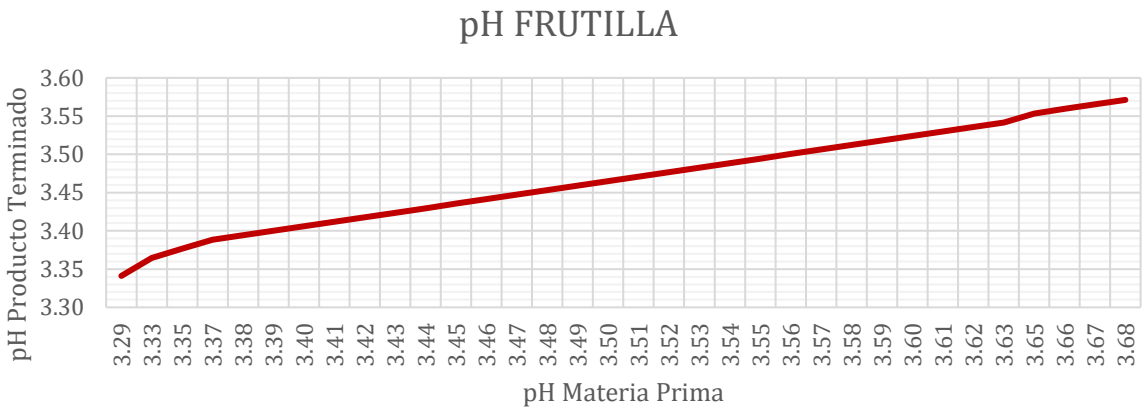
MORA

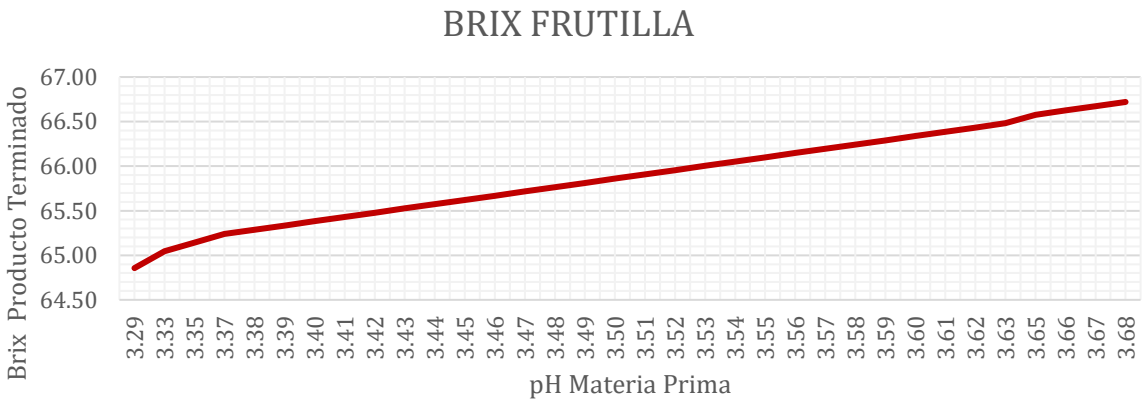
PIÑA



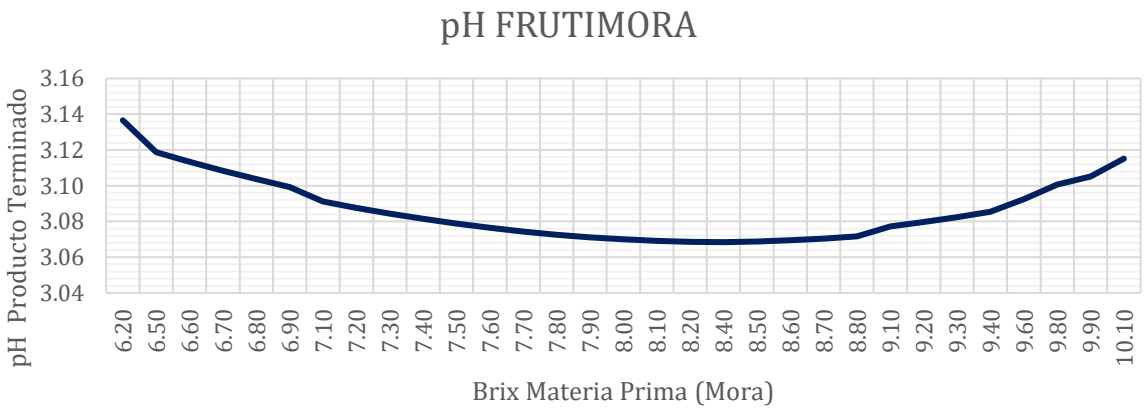
ANEXO V: GRÁFICOS ANÁLISIS DE SENSIBILIDAD

FRUTILLA



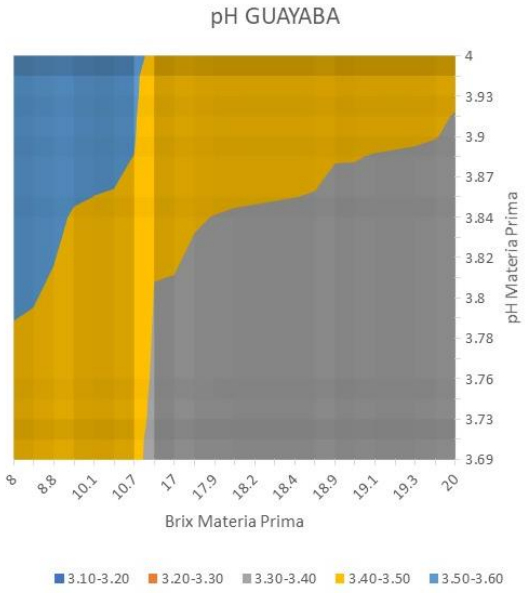


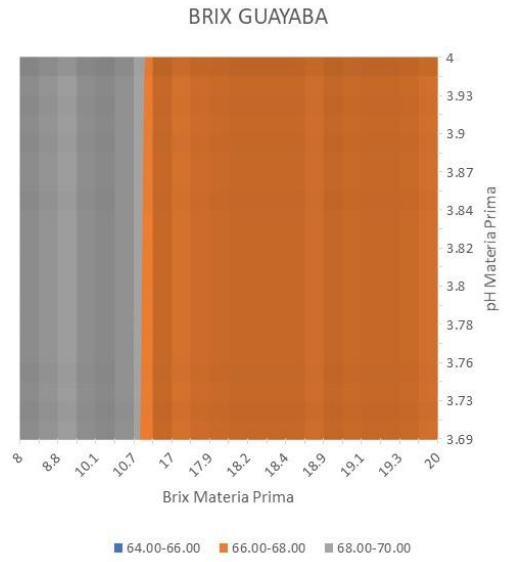
FRUTIMORA



ANEXO V: GRÁFICOS ANÁLISIS DE SENSIBILIDAD (CONT.)

GUAYABA





ANEXO V: GRÁFICOS ANÁLISIS DE SENSIBILIDAD (CONT)

PIÑA