

Evaluación de modelos de aprendizaje estadístico para la
estimación de variables biofísicas en cultivos de caña de azúcar
mediante sensores remotos

Proyecto de Maestría



Camilo Alberto Herrera Rozo

Director

Pablo A. Cipriotti

Ingeniero Agrónomo - Dr. Ciencias Agropecuarias

Co-Director

Héctor A. Chica Ramírez

Agrónomo - Msc. Enseñanza de Matemáticas y Estadística

Biometrista en Cenicaña

Universidad de Buenos Aires

Facultad de Agronomía

Maestría en Biometría y Mejoramiento

Buenos Aires - Argentina, Agosto

2014

1. Introducción

La Agricultura de precisión y el manejo sitio-específico (AEPS), se define como el arte de realizar las prácticas agronómicas requeridas por una especie vegetal, de acuerdo con las condiciones espaciales y temporales del sitio donde se cultiva, para obtener de ellas su rendimiento potencial (Sandoval et al., 2012). La medición de variables biofísicas como área foliar (AF) y el índice de área foliar (IAF) en cultivos de caña de azúcar se realiza en la actualidad por métodos destructivos directos sobre superficies de muestreo pequeñas. El conteo de la cantidad de tallos (NT) en cultivos de caña de azúcar es también un proceso de estimación empírico, no destructivo pero generalmente poco preciso. Estas dos características plantean un panorama interesante de investigación con la finalidad de utilizar nuevos métodos de estimación indirecta, sobre áreas de mayor tamaño (exhaustivos) y de mayor precisión, para los cuales la estadística puede ser una ayuda invaluable.

En tal sentido, existen hoy una gran cantidad de métodos estadísticos que permiten apreciar un abanico muy amplio de posibilidades de análisis, cubriendo cualquier tipo de variable diseño y/o aplicación. El uso de la inteligencia artificial (visión computacional, sistemas expertos, sistemas de ayuda de decisión, máquinas de aprendizaje, etc), y otras técnicas potenciales (redes neuronales, lógica difusa y bioinformática) pueden proporcionar soluciones a distintos problemas en sistemas agrícolas complejos de manera eficaz (Bustos M, 2005). En este contexto, para analizar estructuras de datos complejas derivadas de la agricultura de precisión, tres métodos son de particular interés de los cuales se pretende evaluar el funcionamiento, la adaptación, la eficiencia y la precisión de los mismos, dado que son herramientas potencialmente importantes de aplicar en casos de aprendizaje y clasificación estadística (Mitchell, 1997)(Samuel, 1959). Las metodologías son: PLS o regresión por mínimos cuadrados parciales (De Jong, 1990), random forest o árboles aleatorios (Breiman, 2001) y el aprendizaje por cuantificación vectorial (LVQ) (Kohonen, 1997). Estos tres métodos serán aplicados a datos derivados de imágenes multiespectrales (Hough, 1991) y simulaciones con modelos de cultivos de caña de azúcar.

En este proyecto de tesis se pretende integrar métodos estadísticos y sistemas de visión artificial (Gonzalez y Woods, 1996) a datos captados por sensores remotos (Gutiérrez, 2006) y adicionalmente se desea cubrir la necesidad cada vez mayor de automatizar y mejorar procesos costosos o de características destructivas relacionados a la toma de información, permitiendo hacer un mejor manejo agronómico y apuntando a sistemas de agricultura de precisión y manejo sitio-específico en cultivos de caña de azúcar. Consecuentemente, el objetivo general de este proyecto es el de integrar métodos estadísticos y sistemas de visión artificial o visión por computador, para mejorar la comprensión de procesos naturales, reflejados por cultivos

de caña de azúcar y captados por sensores remotos. Son objetivos específicos los siguientes:

- Evaluar el funcionamiento, adaptación, eficiencia y precisión de tres algoritmos de inteligencia artificial [aprendizaje por cuantificación vectorial (LVQ)], [regresión por mínimos cuadrados parciales] y [random forest] aplicados a datos derivados de imágenes multiespectrales de cultivos de caña de azúcar.
- Estimar el área foliar (AF), el índice de área foliar (IAF) y cantidad de tallos (NT) en cultivos de caña de azúcar mediante algoritmos de inteligencia artificial.
- Ajustar modelos estadísticos y de inteligencia artificial que relacionen correctamente la información derivada de las imágenes multiespectrales a las variables de cultivo (AF), (IAF) y (NT).

2. Revisión de la literatura

La caña de azúcar (*Saccharum officinarum* L. (W.D.Clayton, 1753)) es un cultivo de zonas tropicales y subtropicales que se propaga mediante la plantación de trozos y/o vástagos de caña, donde de cada nudo sale una planta nueva e idéntica a la original. Existen una infinidad de métodos para la evaluación del crecimiento y desarrollo de los cultivos, entre estos métodos encontramos nuevas aplicaciones como la percepción remota o teledetección, la cual se define como el grupo de técnicas para la obtención de información sobre las propiedades físicas de ciertas superficies u objetos y su entorno, desde distancias relativamente grandes, sin contacto físico con ellos (Denniss, 1995). Las imágenes adquiridas por sensores desde plataformas aéreas o satelitales en el mundo agropecuario tienen un potencial que se ha venido explorando con mayor énfasis en la última década (Gutiérrez, 2006).

El IAF es el cociente entre el área foliar (AF) y la unidad de superficie de suelo. La información precisa y oportuna sobre el IAF tiene gran importancia y aplicaciones en la agricultura para la estimación del rendimiento y la evaluación de estrés en distintos cultivos, y en la ecología para el estudio de la producción primaria y el cambio ambiental (Curran y Steven, 1983). Las principales aplicaciones de las técnicas de teledetección están dentro de los campos de la inteligencia agrícola, la gestión agrícola y la investigación ecológica (Curran y Steven, 1983). Desde los inicios de la percepción remota los índices espectrales de vegetación han sido útiles y fáciles de calcular para relacionarlos con diversas variables agronómicas. Aunque los índices espectrales de vegetación en muchos casos muestran excelentes relaciones con estas variables, es necesario calibrar o comprender la equivalencia de sus valores en la estimación de contenidos de clorofila, IAF o biomasa (Sandoval et al., 2012).

Hoy en día existen muchas técnicas estadísticas que el agricultor puede aprovechar para la estimación de variables del cultivo, entre estas está el aprendizaje automático que es una rama de la Inteligencia Artificial (IA) cuyo objetivo es desarrollar técnicas que permitan a las computadoras aprender. Una máquina es un sistema organizado capaz de transformar un cierto mensaje de entrada en otro de salida, de acuerdo con algún principio de transformación. Si tal principio está sujeto a cierto criterio de validez de funcionamiento, y si el método de transformación se ajusta a fin de que tienda a mejorar el funcionamiento del sistema de acuerdo con este criterio, se dice que el sistema aprende.

En (Samuel, 1959) se define a las máquinas de aprendizaje o el aprendizaje automático como el campo de estudio que da a las computadoras la capacidad de aprender sin ser explícitamente programadas. Esto sería un gran avance en las técnicas de estimación en cultivos de caña de azúcar, sobre todo cuando la información proviene de imágenes digitales, proceso que genera un flujo grande de información, el cual requiere métodos estadísticos capaces de tratar con estructuras de datos multivariados y no lineales. Las herramientas matemáticas basadas en la minería de datos proporcionan un marco adecuado para la extracción de información útil a partir de grandes bases de datos, así como también pueden conducir al descubrimiento de conocimiento (Ferraro, 2008).

La regresión por mínimos cuadrados parciales o PLS, el algoritmo de random forest o árboles aleatorios y el aprendizaje por cuantificación vectorial (LVQ) son tres metodologías de mucho interés para el desarrollo de esta investigación pues aunque son métodos cada uno desarrollado para aplicaciones originales diferentes, a priori son altamente escalables y se espera obtener resultados originales con la aplicación de métodos que con anterioridad no se han probado sobre cultivos de caña.

2.1. Regresión por mínimos cuadrados parciales (PLS)

La regresión por mínimos cuadrados parciales se introdujo hace casi treinta años y ha tenido un gran desarrollo en áreas como la quimiometría, donde se analizan datos que se caracterizan por muchas variables predictoras, con problemas de multicolinealidad, y pocas unidades experimentales en estudio (Geladi y Kowalski (1986) , De Jong (1990)).

Es una forma particular de análisis multivariante, relacionado con la regresión de componentes principales (PCR) y posee valiosas ventajas teóricas y computacionales que han llevado a innumerables aplicaciones. PLS se utiliza para encontrar las relaciones fundamentales entre dos matrices (X e Y), es decir, un enfoque de variable latente para el modelado de las estructuras de co-varianza en estos dos espacios.

Un modelo de PLS trata de encontrar el sentido multidimensional en el espacio X que explica la dirección de la máxima varianza multidimensional en el espacio Y (Tenenhaus et al., 2005). Estos métodos tienen ventajas intrínsecas cuando se los compara con métodos univariados, Todas las variables relevantes son incluidas en el modelo PLS. La suposición básica de todos estos modelos es que el sistema o proceso estudiado depende de un número pequeño de variables latentes (V.L.). Este concepto es similar al de componentes principales. Las variables latentes son estimadas como combinaciones lineales de las variables observadas.

2.2. Árboles Aleatorios

Breiman (1996, 2000, 2001, 2004) demostró que hay una ganancia sustancial en la precisión en los métodos de clasificación y regresión mediante el uso de un conjunto de árboles donde cada árbol en el conjunto se cultiva de acuerdo a un parámetro aleatorio. Las predicciones finales se obtienen de las agregaciones sobre el conjunto de datos. Como los componentes de base del conjunto son predictores con estructura de árbol, y desde cada uno de estos árboles se construye utilizando una introducción de aleatoriedad, se conoce a estos procedimientos como como árboles aleatorios.

El random forest es una combinación de árboles predictores tal que cada árbol depende de los valores de un vector aleatorio probado independientemente y con la misma distribución para cada uno de éstos. Es una modificación sustancial de bagging [Bootstrap aggregating] que construye una larga colección de árboles no correlacionados y luego los promedia. Random forest es una unión entre métodos de clasificación y métodos de regresión, esta opera por la construcción de una multitud de árboles de decisión en el entrenamiento y en la salida una clase el cual es el modo de salida de los árboles individuales.

2.3. Aprendizaje por cuantificación vectorial (LVQ)

LVQ se puede entender como un caso especial de una red neuronal artificial, con mayor precisión, este aplica el concepto de “el ganador toma todo” mediante el aprendizaje de Hebb(Hebb, 1949), es un precursor de los mapas auto-organizados (SOM) y está relacionado con el algoritmo de gas neural y con el algoritmo de K vecinos más cercanos (k-NN) (Kohonen, 1997).

El aprendizaje por cuantificación vectorial (LVQ) supone una extensión del aprendizaje competitivo donde los prototipos están etiquetados. Ahora, además de considerar la cercanía de un prototipo se puede evaluar la clase de éste e imponer, por lo tanto, correcciones de premio (acercamiento) o castigo (alejamiento). En ciencias de la computación, el aprendizaje

por cuantificación vectorial (LVQ), está basado en prototipos de algoritmos de clasificación supervisada. LVQ es la contraparte supervisada de los sistemas de cuantificación vectorial (Hastie et al., 2009).

2.4. Validación cruzada

La validación cruzada o cross-validation es una técnica utilizada para evaluar los resultados de un análisis estadístico y garantizar que son independientes de la partición entre datos de entrenamiento y prueba. Consiste en repetir y calcular la media aritmética obtenida de las medidas de evaluación sobre diferentes particiones. Se utiliza en entornos donde el objetivo principal es la predicción y se quiere estimar qué tan preciso es un modelo que se llevará a cabo a la práctica (Devyver y Kittler, 1982).

Es una técnica muy utilizada en proyectos de inteligencia artificial para validar modelos generados. Probablemente el método más simple y ampliamente usado para estimar el error en predicción es la validación cruzada, este método estima directamente el error esperado de una muestra-extra $Err = E \left[L \left(Y, \hat{f}(X) \right) \right]$, el error medio generalizado cuando el método $\hat{f}(X)$, se aplica a una muestra de prueba independiente de la distribución conjunta de X e Y. Como se mencionó anteriormente, se podría esperar que la validación cruzada estime el error condicional, con el conjunto de entrenamiento τ el cual se mantiene fijo (Hastie et al., 2009).

2.5. Validación cruzada K-veces (“K-Fold”)

Sería ideal que, en caso que se tuvieran datos suficientes, se dejara de lado un conjunto de datos para la validación y los restantes se usaran para evaluar el desempeño de nuestro modelo de predicción. En general no es posible dado que los datos son a menudo escasos, y la precisión de la validación es perjudicada. Alternativamente, si se repite el procedimiento en K ocasiones, cada una con datos distintos para entrenamiento y validación, se gana en precisión. Para ello se dividen los datos en K partes más o menos del mismo tamaño; por ejemplo, cuando $K = 5$, el escenario podría ser como sigue:

1	2	3	4	5
Train	Train	Validation	Train	Train

Figura 1: k-fold

Para el grafico anterior en $k=3$, se ajusta el modelo con las otras $k-1$ partes de los datos, es decir con ($k=1$, $k=2$, $k=4$, $k=5$) y se calcula el error de predicción del modelo para la k -esima parte, esto para predecir la parte de orden k (en este caso la parte 3) de los datos, se realiza este proceso para cada $k=1,2,3,...,K$ y se combinan las estimaciones del error de predicción.

Ahora sea $\kappa : \{1, \dots, N\} \rightarrow \{1, \dots, K\}$ una función de indexación la cual que indica a que partición es asignada la observación i por la aleatorización. Denotemos por $\hat{f}^{-k}(X)$ la función de ajuste, computada con la k -esima parte de los datos eliminada, entonces la validación cruzada estimada del error de predicción es:

$$CV(\hat{f}) = \frac{1}{N} \sum_{i=1}^N L(y_i, \hat{f}^{-\kappa(i)}(x_i))$$

La típica elección es $k = 5$ o 10 , pero incluso se podría tener una elección de $K=N$ lo que es conocido como una validación cruzada dejando solo uno fuera “leave-one-out ” y donde el estimador de validación cruzada es aproximadamente insesgado, pero no se utiliza mucho este caso ya que cuando tenemos tantos conjuntos de validación, existen muchos conjuntos casi iguales. Por esta razón en la mayoría de las veces se opta por selección de un K un poco más conservador como 5 o 10 , lo que es suficiente para obtener buenos resultados (Hastie et al., 2009).

2.6. Medidas de Ajuste

En el campo de las máquinas de aprendizaje, el rendimiento de un clasificador se mide generalmente en términos de error de predicción. En la mayoría de los problemas del mundo real, el error no se puede calcular con exactitud y debe ser estimado. Por lo tanto, es importante elegir un estimador apropiado del error (Rodriguez, 2010).

En general:

Positivo (P) = identificado Negativo (N) = rechazado

Por lo tanto:

Verdadero Positivo (VP)= identificado correctamente

Verdadero Negativo (VN)= rechazado correctamente

Falso Positivo (FP) = identificado incorrectamente

Falso negativo (FN) = rechazado incorrectamente

En estadística, el error cuadrático medio o MSE (Mean squared error) de un estimador mide el promedio de los errores al cuadrado, es decir, la diferencia entre el estimador y lo que se estima. El MSE es una función de riesgo, correspondiente al valor esperado de la pérdida del error al cuadrado o pérdida cuadrática. La diferencia se produce debido a la aleatoriedad o porque el estimador no tiene en cuenta la información que podría producir una estimación más precisa (Lehmann y Casella, 1998).

Error cuadrático medio (MSE): $\frac{1}{n} \sum_{i=1}^n (Predicción_i - DatosReales_i)^2$

En una analogía con la desviación estándar, tomando la raíz cuadrada del MSE produce el error de la raíz cuadrada de la media o la desviación de la raíz cuadrada media (RMSE o RMSD), que tiene las mismas unidades que la cantidad que se estima; para un estimador insesgado, el RMSE es la raíz cuadrada de la varianza, conocida como la desviación estándar.

Raíz del error cuadrático medio (RMSE): $\sqrt{\frac{1}{n} \sum_{i=1}^n (Predicción_i - DatosReales_i)^2}$

Algunos nombres que adquieren características relacionadas a la medición del error de predicción:

1. Error cuadrático medio (o la raíz de error cuadrático medio)
 - Sensible a outliers
2. Desviación media absoluta
 - A menudo más robustos
3. Sensibilidad - Sensitivity (recall)
 - Proporción de positivos correctamente identificados, Si se quieren pocos positivos perdidos es de gran interés controlarla.
4. Especificidad - (Specificity)
 - Es llamada a la capacidad de nuestro estimador para dar como casos negativos los casos realmente negativos, Si se quieren pocos negativos llamados como positivos es de gran interés controlarla.
5. Precisión (Accuracy)

- Es el resultado de la realización de una curva ROC la cual es una representación gráfica de la sensibilidad frente a $(1 - \text{especificidad})$ para un sistema clasificador binario según se varía el umbral de discriminación.
- La precisión está definida como $\frac{(VP+VN)}{(P+N)}$ y se interpreta como el equilibrio de los pesos de los falsos positivos y falsos negativos.

6. Concordancia

- Se dice que un instrumento o procedimiento es preciso si sus resultados son consistentes cuando se aplica más de una vez al mismo individuo bajo las mismas circunstancias, El Coeficiente kappa de Cohen mide el acuerdo entre dos evaluadores que cada uno clasifica los elementos de N en C categorías mutuamente excluyentes. La primera mención de una estadística kappa como se atribuye a (Galton, 1892).

3. Antecedentes

Hace más de 60 años que se toman fotografías a color e infrarrojo para seguir el crecimiento de plantas Colwell (1956). En la actualidad, estos métodos están siendo reevaluados para realizar análisis dentro de la variabilidad espacial en la agricultura de precisión, ya que las imágenes aéreas se pueden adquirir rápidamente durante los períodos críticos del crecimiento de las plantas (Blackmer et al., 1996).

A nivel mundial países como Francia y Brasil han realizado trabajos para estimar algunos parámetros biofísicos e inclusive para pronosticar la producción de caña de azúcar. En Francia Bégué et al. (2008) haciendo uso del índice de vegetación normalizado (NDVI) en cultivos con variabilidad espacial (independiente a las etapas de crecimiento de los cultivos), demostraron que en una escala estacional, el patrón de crecimiento dentro de un campo depende de la etapa fenológica del cultivo mientras que a escala anual los mapas NDVI revelaron patrones estables. Lo anterior permite concluir que es necesario conocer el ciclo de crecimiento del cultivo para interpretar correctamente los patrones espaciales. Las imágenes de una fecha única pueden ser insuficientes para el diagnóstico de la situación de los cultivos o para aplicaciones en predicción.

En Sandhu et al. (2012) se evaluaron las calificaciones visuales subjetivas de crecimiento del cultivo de la caña de azúcar para la estimación de parámetros de rendimiento del cultivo y paralelamente se realizaron mediciones de los IAF. En este caso se encontró que existe relaciones importantes entre estos sistemas de evaluación visual, los índices IAF y la estimación

de la población, pero estas relaciones son válidas solo en algunos estados del crecimiento del cultivo. Por lo tanto, las estimaciones no eran buenas fuera de algunos periodos del desarrollo del cultivo, sobre todo en las primeras etapas del crecimiento, lo que impide tomar decisiones tempranas respecto al manejo del cultivo.

En Brasil Almeidaa et al. (2006) ha propuesto un método para realizar predicción del rendimiento sobre cultivos de caña de azúcar usando índices de vegetación espectral, mediante análisis de componentes principales e información histórica de los cultivos. Para dicho estudio se utilizaron imágenes (ETM +) / Landsat-7 e imágenes ASTER/Terra. Este método comprende varias etapas y permite una síntesis de toda la información tanto de la imagen como de la información histórica, normalizando todas las variables en conjunto, y haciendo posible expresar todos los datos en imágenes síntesis.

Sudáfrica es el líder en producción de caña de azúcar en África y uno de los más grandes productores del mundo, El monitoreo de estos factores de estrés del cultivo de caña es de vital importancia para tomar acciones preventivas y de mitigación sobre el cultivo. Abdel-Rahman (2010) exploró el potencial de usar sensores remotos en cultivos de caña de azúcar, mediante el uso de imágenes Landsat TM y ETM+ con las cuales generó modelos de predicción de rendimiento de caña aplicando algoritmos de random forest optimizados, logrando una estimación buena para algunas de las variedades estudiadas.

4. Metodología

El desarrollo de este proyecto de investigación se realizará en cultivos experimentales de caña de azúcar los cuales se encuentran ubicados en el valle geográfico del río Cauca, Colombia. Dichos cultivos experimentales se utilizarán por la necesidad de realizar tomas de información sobre cultivos donde se puedan controlar las distintas fuentes de variación como sistemas de riego, el tamaño del lote productivo, la fecha y densidad de siembra, el manejo del suelo, el seguimiento del cultivo durante el crecimiento, y su fertilización, entre otros factores. Así, se limitará la variabilidad para realizar una correcta calibración y aproximación de las variables a estimar.

Estos cultivos son campos de investigación de [CENICAÑA - Centro de Investigación de la Caña de Azúcar de Colombia] que han sido designados para este fin, los cuales poseen unas características específicas como el tipo de suelo en el que se encuentran, además de las características climáticas e inclusive en muchos casos se tienen registros históricos de su producción. Respecto a los tipos de suelos y características del área de estudio, los cultivos evaluados estarán sobre áreas agroecológicas catalogadas como **6H1** (Suelos de texturas finas y contenidos de arcilla entre 35 % y 60 %, moderadamente bien drenados que poseen características de permeabilidad baja < 200 mm/año) área predominante en el Valle del Cauca, de acuerdo a la clasificación presentada en (Carbonell G. J., 2001) y se evaluarán las variedades genéticas de caña CC-8592 y CC-934418, las cuales son de gran interés de estudio para el centro de investigación.

Una vez realizadas las pruebas de estimación sobre los cultivos, se tendrá una metodología definida para la estimación de las variables biofísicas y del tamaño la población del cultivo partir de imágenes multiespectrales de alta resolución. El procedimiento de toma de información se realizará mediante aviones ultra-livianos de fumigación del cultivo con una cámara multi-espectral ¹ (Figura: 2) para toma de imágenes en campo enlazada a sistemas de GPS, que permitirán una mejor geolocalización de las tomas (imágenes) realizadas.

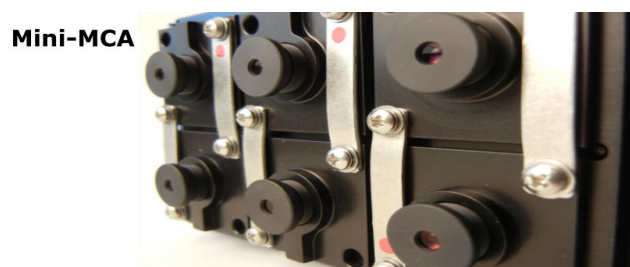


Figura 2: Cámara Multi-espectral

¹http://www.tetracam.com/Products-Mini_MCA.htm

Se estudiará la información reflejada por el cultivo mediante imágenes multiespectrales en formato raster (imagen matricial), donde la información de cada imagen se considera como una matriz de datos de dimensiones ($n \times m$), siendo n el número de píxel verticales y m el número de píxel horizontales. Estas dimensiones están determinadas por la resolución de la cámara mediante la cual se realiza la toma de la información. Además en este caso al tratarse de imágenes multiespectrales cada imagen cuenta con seis bandas espectrales las cuales corresponden a un rango del espectro.

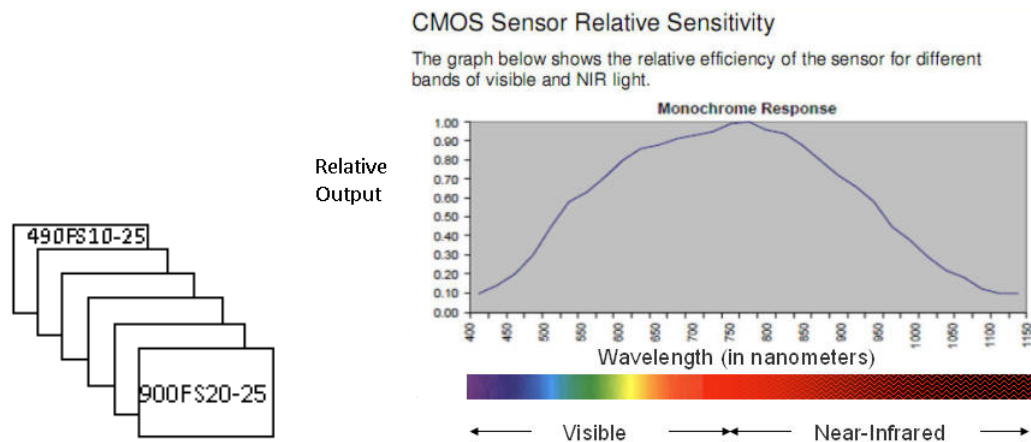


Figura 3: Rango Espectral - mini-MCA

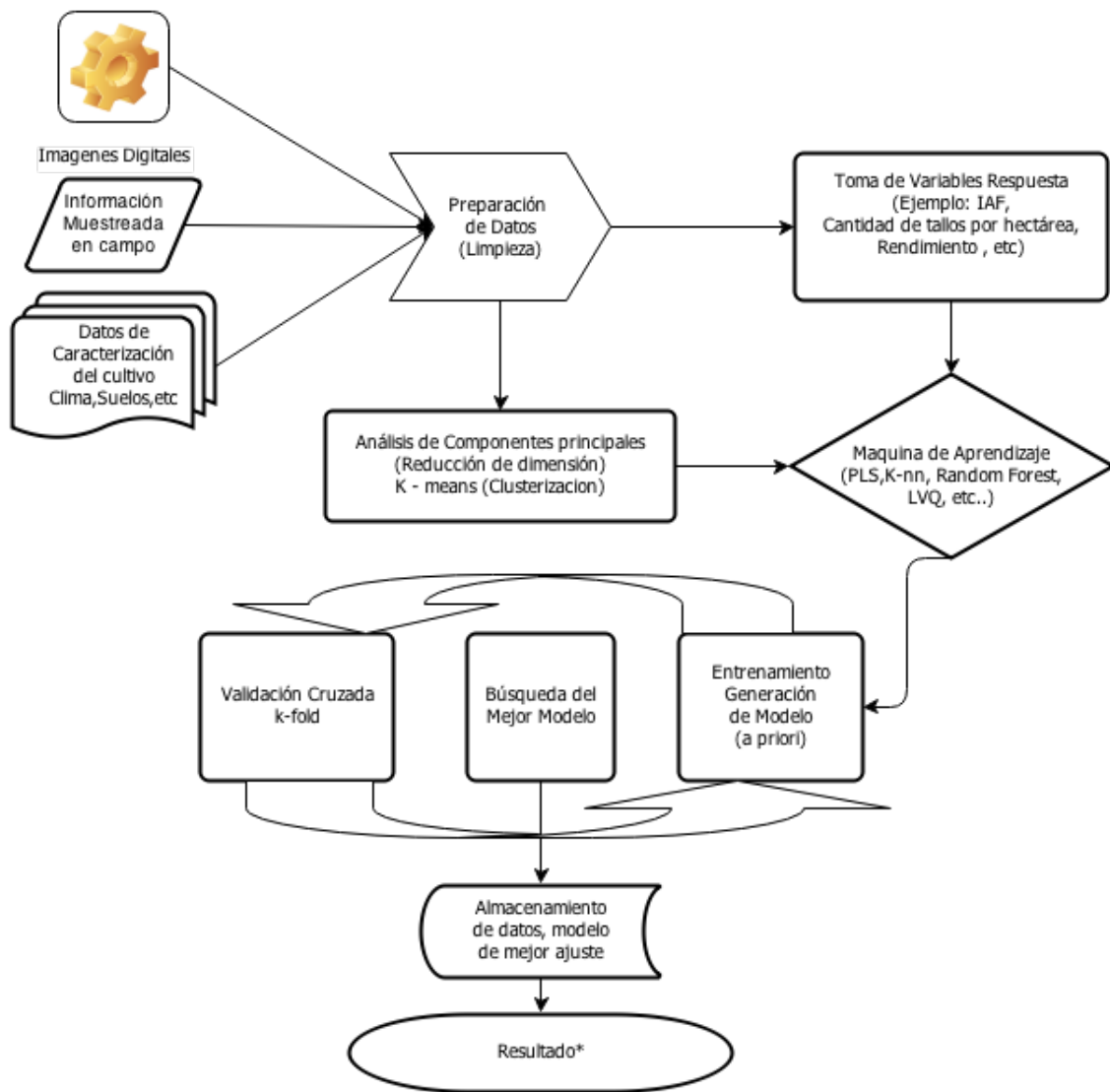
Para el caso específico de la cámara tetracam mini-MCA de seis lentes las bandas espectrales para cada lente están definidas así: 490FS10-25 , 550FS10-25, 680FS10-25, 720FS10-25, 800FS10-25 , 900FS20-25, con esta configuración de lentes se abarca la mayoría del espectro visible e infrarrojo (Figura: 3). Cada imagen debe ser corregida de deformaciones y sobrepuesta en un SIG que delimitará el área de estudio, así se evitará que ingrese información adicional que dificulte el correcto análisis de información contenida en los datos, esto permitirá realizar un análisis sobre cada lote productivo.

En campo se tomarán muestras directas en los cultivos evaluados, con la finalidad de realizar una calibración del procedimiento de estimación. Esta calibración se llevará a cabo mediante la toma de información de caracterización de suelos y principalmente de variables biofísicas y de población que se desea estimar. Se realizará un conteo manual del número de tallos mientras que en áreas pequeñas al interior del campo se realizará una muestra (destruccion) para la medición de las variables biofísicas (AF , IAF). Así, con esta información de muestreo será posible realizar un proceso de validación cruzada junto a las imágenes tomadas del campo lo que permitiría mejorar mucho más las estimaciones finales realizadas a partir de las imágenes.

Lograr estimar la cantidad de toneladas de caña por hectárea de manera temprana es uno de los objetivos a nivel industrial. En la teoría es posible estimar estas cantidades haciendo uso de las variables biofísicas y de población antes señaladas, pero en la práctica no se realiza, ya que algunas de las variables biofísicas son medidas mediante tomas de información de carácter destructivo y con costos elevados por lo cual no es posible hoy día.

En la actualidad está cambiando el paradigma de cómo obtener la información a como almacenarla y analizarla adecuadamente. Anteriormente la información se recolectaba mediante métodos tradicionales que no entregaban volúmenes de datos muy grandes. Ahora se posee una gran cantidad de sensores que pueden recolectar una cantidad inmensa de información, la cual será la que alimentará los modelos para finalmente realizar una aproximación (modelación) a los sucesos o procesos naturales involucrados, esto con la necesidad de maximizar los beneficios de dichos datos capturados. Dado este panorama hoy en día existe una gama más amplia de algoritmos y métodos estadísticos que permiten realizar un tratamiento y modelado de información más amplio (Mitchell, 1997).

Es de gran importancia poder elegir con claridad cuál es el método más adecuado en pro de la calidad de los resultados obtenidos, del procesamiento y modelado de la información, por lo cual se presenta una metodología general donde se nombran técnicas de gran interés que podrán ser evaluadas en el proceso de esta investigación. A continuación se ve una propuesta analítica básica de cómo se pretende abordar el problema de investigación.



*Generación de metodología de clasificación y aproximación a la respuesta en el Área Foliar e IAF y cantidad de tallos por lote productivo de caña de azúcar, determinado por imágenes Multiespectrales y variables espacio temporales.

Figura 4: Propuesta Analítica Básica

5. Flujo de procesos:

El insumo inicial para el desarrollo de este proyecto son las imágenes multiespectrales tomadas sobre los cultivos de caña que se desean evaluar, estas imágenes deben ser georeferenciadas y puestas sobre un SIG una vez estas imágenes tienen estas características se procederá a realizar los siguientes procesos sobre la información disponible:

Primero vamos a definir los sets de datos involucrados en el proceso y qué características están relacionadas a dichos datos:

1. Variables iniciales

- Usar los valores de las bandas espectrales de las imágenes como variables.
- Generar índices espectrales a partir de las bandas espectrales para agregarlos como variables.
- Recopilar información adicional tomada al principio del proyecto en campo por ejemplo: variables relacionadas a suelos y otras variables medidas por muestreo que caractericen el área de estudio.

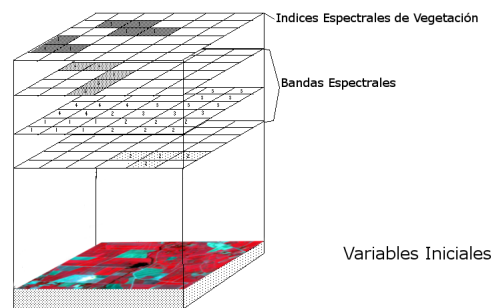


Figura 5: Variables Iniciales

2. Variables respuesta

- Recopilar información de área foliar e IAF por muestreo
- Recopilar información de cantidad de tallos por lote productivo

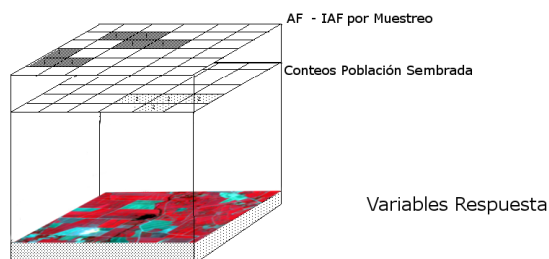


Figura 6: Variables Respuesta

Con la información de las variables iniciales se realizarán los siguientes procesos:

- Estandarizar toda la información al tamaño de píxel adecuado para maximizar la eficiencia del proceso y manejar la misma resolución espacial en todo el proceso.
- Se realizará un proceso de reducción de dimensionalidad considerando distintos sets de variables [Mediante Análisis de Componentes Principales (Shlens, 2005)].
- Realizar un proceso de clusterización o agrupamiento de píxeles con características de respuesta espectral ajustando píxeles similares a categoría encontradas en los datos, donde dentro de cada clase se minimiza la variabilidad y entre clases se maximiza la diferencia entre estos. [Algoritmo K-medias (Hartigan, 1975)].

Con la información denominada variables respuesta se realizará el siguiente proceso:

- Estandarizar toda la información al tamaño de píxel o resolución a igual escala a la tomada en los datos de variables iniciales.

Después de realizar estos procesos previos existen dos posibilidades para abordar el problema:

- Usar solo los píxeles que tienen información tanto para las variables iniciales como para las variables respuesta (dependerá de la cantidad de muestras tomados sobre el lote productivo) , con esta información entrenar sistemas de aprendizaje o modelos estadísticos.
- Realizar una interpolación en las variables respuesta a píxeles sin información que contengan características de similar comportamiento en la clasificación de los datos iniciales (no dependerá necesariamente de la cantidad de muestras tomados sobre el lote productivo), con esta información entrenar sistemas de aprendizaje o modelos estadísticos.

Cualquiera de los dos procesos tomados tendrá como resultado un modelo de aproximación a la respuesta de las variables biofísicas o de población que se pretenden estimar, por lo cual es de gran importancia escoger un adecuado proceso que permita registrar las estructuras de interacción compleja en los datos de una forma adecuada.

La regresión por mínimos cuadrados parciales (PLS), los árboles aleatorios y el aprendizaje por cuantificación vectorial (LVQ), son metodologías de un gran interés ya que una potencial aplicabilidad desde la teoría (Breiman, 2001);(Hastie et al., 2009);(Kohonen, 1997), pero se

hará uso de aquel método que entregue mejores resultados respecto de las variables a estimar AF , IAF y NT, y se validará mediante un proceso de validación cruzada.

Se evaluará el funcionamiento de cada uno de estos algoritmos mediante la puesta en práctica de éstos sobre datos derivados de imágenes multiespectrales. Este funcionamiento se podrá ver y evaluar siempre que se dé la posibilidad de adaptar estos métodos nombrados a esta nueva metodología desarrollada explícitamente para resolver los problemas de estimación de variables biofísicas. Por eso es de principal interés probar que los conceptos teóricos detrás de estos modelos son aplicables a nuestros datos. Posteriormente, se evaluará qué tan eficiente es cada algoritmo tanto lo referente al desarrollo como a tiempos de ejecución, ya que al considerar estructuras de datos complejas tanto multivariantes como espaciales hay algoritmos que pueden asimilar mejor o peor estas estructuras. Finalmente, se medirá la sensibilidad, especificidad y precisión de él o de los modelos que resulten de los procesos previos.

Una vez se han evaluado todas estas alternativas, se seleccionará el método que permita realizar una mejor estimación de las variables biofísicas y de densidad poblacional o en su defecto una combinación de métodos que mejore los resultados esperados. Esta selección se realizará mediante un proceso de validación cruzada el cual consta de dos subsets del total de los datos, donde uno de los subsets se utiliza para entrenamiento del modelo evaluado y el otro para la validación del resultado obtenido mediante los datos derivados de las imágenes multiespectrales en términos de la respuesta observada.

6. Consideraciones respecto a la cantidad de información a ser procesada en esta investigación:

Se considerará una cantidad de imágenes de lotes productivos, que sea efectivamente muestreable tanto mediante la toma de imágenes como a partir del personal en campo, además de considerar la capacidad de análisis de la información muestreada. Por último, se evaluará la resolución espacial final (tamaño del pixel) mediante la cual se realizará el análisis de los datos, esto se definiría intentando lograr un balance entre la ganancia en resolución espacial y la capacidad computacional para realizar el procesamiento de los datos. Esto con el propósito de asegurar el pos-procesamiento de toda la información y extraer el máximo beneficio de los datos adquiridos. Además, dependiendo de la resolución espacial lograda en las imágenes, es posible explorar metodologías de segmentación artificial sobre estas imágenes, que permita una mejor visualización de las estructuras espaciales intrínsecas presentes en los cultivos.

7. Justificación

En Colombia los costos para producir una tonelada de azúcar son mayores que en otros países, a causa de múltiples factores como el precio de la tierra, el costo de la mano de obra, los insumos, los impuestos, la seguridad, el valor de los derechos de uso del agua, el costo de fertilizantes e inclusive los costos de inspecciones para la detección de plagas y enfermedades (Sandoval et al., 2012). Igualmente, los recursos necesarios para medir la sacarosa en la caña en evaluaciones de pre cosecha, y otros rubros de importancia. Una manera de disminuir estos costos de producción y mejorar la productividad de los campos es mediante la utilización de herramientas para identificar tempranamente las condiciones anormales en los cultivos, de manera que se pueden tomar medidas de control o de prevención oportunas. El uso de la percepción remota en combinación con sistemas de información geográfica (SIG) y sistemas de posicionamiento global (GPS), generan una oportunidad para mejorar la competitividad de la producción agrícola asegurando el desarrollo sostenible de la actividad.

Las variables que se desean estimar por medio de esta metodología propuesta son de vital importancia para poder efectuar pronósticos tempranos de producción y sistemas de manejo sitio específico en el cultivo caña de azúcar entre otras aplicaciones, lo que generaría avances importantes en el sector agro industrial de caña de azúcar en el Valle del Cauca relacionados a los siguientes factores:

1. Innovación

- Actualmente no se utiliza fotografía aérea para conocer el estado del cultivo y tampoco se ha desarrollado algoritmos de inteligencia artificial para vincular imágenes con la estimación de parámetros biofísicos del cultivo.

2. Utilidad Económica

- Actualmente se utilizan métodos de medición sobre cultivos de caña de azúcar que conllevan altos costos económicos y logísticos.

3. Facilidad de aplicación

- El sensor puede instalarse en plataformas aéreas, como aviones de fumigación, UAVs, grúas u otros y la información puede ser procesada fácilmente una vez se establezca un protocolo estándar resultante de esta investigación.

4. Impacto ambiental

- Conocer el comportamiento del cultivo en una etapa temprana (4 meses - 6 meses) donde se expresan características como su crecimiento y desarrollo, permitiría realizar tanto estimaciones tempranas de producción como también evaluar el estado del cultivo. Esto permitiría realizar acciones de corrección cuando estas son necesarias, logrando minimizar el impacto en costos y sobre el suelo, que se podrían producir si estas acciones se toman en una etapa posterior al periodo de desarrollo del cultivo.

Los ingenios tienen la necesidad de cuantificar el rendimiento de los cultivos de caña de azúcar, ya sea como un método de seguimiento del cultivo. Al lograr mejorar la resolución de la toma de información se enriquece mucho más el conocimiento del cultivo y así puede llevar a tener en cuenta prácticas realizadas sobre el cultivo que harían más preciso y menos empírico el manejo de los cultivos en pro del rendimiento y calidad en la producción.

Realizar un proceso adecuado de estimación de parámetros biofísicos en el cultivo de caña a partir de imágenes multiespectrales genera una gran ganancia al estudio del comportamiento de la caña a nivel de cultivo, además de proveer información que hasta el momento solo es capturada mediante costosos instrumentos, muchas veces necesariamente estos procesos de medición deben ser destructivos. Además logrando capturar una mejor resolución espacial en imágenes multiespectrales junto con SIG es posible generar mapas de mayor nivel de detalle para características importantes tanto del cultivo como de los suelos, dándole un valor agregado a cada área estudiada.

8. Significado de la investigación

Se desea mostrar que el proceso de estimación de características, tanto de la población (número de tallos) como biofísicas, (IAF) , presente en cultivos de caña es posible por medio del uso de imágenes multiespectrales y un correcto procesamiento estadístico de los datos mediante algoritmos no explorados respecto a su aplicación en cultivos de caña de azúcar. Además se espera lograr la aplicabilidad real de este proceso de estimación para medir de forma no destructiva las variables biofísicas, sobre cultivos de caña que de otra manera no es posible. Adicionalmente, se analizará la factibilidad de implementar nuevas estrategias de modelación del cultivo y captura de datos para aprovechar al máximo toda la información presente en el cultivo.

Facilidades Disponibles

Este proyecto hace parte fundamental de un proyecto global en desarrollo por el programa de agronomía y el área de geomática del Centro de Investigación de la Caña de Azúcar de Colombia (CENICAÑA) titulado “Evaluación de imágenes de alta resolución para la estimación de parámetros biofísicos en caña de azúcar” por lo cual cuenta con el apoyo e interés del centro de investigación para su desarrollo y consolidación.

Los datos serán adquiridos con los siguientes insumos disponibles para el desarrollo de este proyecto: para la toma de información aérea, se cuenta con un sistema de cámara multi-espectral tetracam mini-MCA de seis lentes, y está desarrollado un protocolo de toma de información mediante aviones de fumigación, dicho protocolo cuenta con la implementación de un sistema de GPS y un sistema inercial para lograr ortogonalidad y precisión en la toma de información. Existe una serie de cultivos designados para realizar la toma de las imágenes. Adicional a esto se encuentra presupuestado el costo de jornales para toma de medidas y métodos destructivos en campo (muestreo).

También se cuenta con el apoyo técnico y logístico de profesionales en sensores remotos, sistemas de información geográfica y biometría vinculados al desarrollo de este proyecto, los cuales están representados por el área de geomática de sistemas de información geográfica y de biometría de la entidad (CENICAÑA). El procesamiento estadístico de la información se realizará por medio de software libre R (Team, 2014) el cual es un excelente elemento computacional, que proporciona un sistema muy versátil y permite que la información sea manejada de una manera idónea. Además, la implementación de rutinas es una herramienta muy útil, que sumada a la gran cantidad de aportes en cuanto a metodologías realizadas por la comunidad, hacen que este sistema sea una herramienta ideal para el desarrollo e implementación de nuevas rutinas que pueden ser desarrolladas de manera amplia y sin mayores limitaciones enfocadas a la estadística y al manejo de la información.

Los gastos adicionales distintos a la toma de información como (un computador (Samsung serie 5 con 12gb de ram), papelería, etc), ya se encuentran disponibles y serán solventados por el investigador.

Bibliografía

- Abdel-Rahman, E. (2010). *The Potential for Using Remote Sensing to Quantify Stress in and Predict Yield of Sugarcane (Saccharum Spp. Hybrid)*. University of KwaZulu-Natal, Pietermaritzburg.
- Almeidaa, T. I. R., Filhob, C. R. D. S., and Rossettoc, R. (2006). Aster and landsat etm+ images applied to sugarcane yield forecast. *International Journal of Remote Sensing*, 27(19):4057–4069.
- Bégué, A., Todoroff, P., and Pater, J. (2008). Multi-time scale analysis of sugarcane within-field variability: improved crop diagnosis using satellite time series. *Precision Agriculture*, 9(3):161–171.
- Blackmer, T. M., Schepers, J. S., Varvel, G. E., and Meyer, G. E. (1996). Analysis of aerial photography for nitrogen stress within corn fields. *Agronomy Journal*, 88(5):729–733.
- Breiman, L. (1996). Bagging predictors. *Machine Learning*, 24(2):123–140.
- Breiman, L. (2000). Some infinity theory for predictor ensembles. Technical Report 577, UC Berkeley.
- Breiman, L. (2001). Random forests. *Machine Learning*, 45(1):5–32.
- Breiman, L. (2004). Consistency for a simple model of random forests. Technical report, UC Berkeley.
- Bustos M, J. R. (2005). Inteligencia artificial en el sector agropecuario. *Seminario Investigativo*, Universidad Nacional de Colombia. Bogotá, Colombia,.
- Carbonell G. J., AMAYA E. A., O. B. e. a. (2001). Zonificación agroecológica para el cultivo de caña de azúcar en el valle del río cauca. cuarta aproximación. Technical Report 29, Cenicaña.
- Colwell, R. (1956). *Determining the Prevalence of Certain Cereal Crop Diseases by Means of Aerial Photography*. [Hilgardia. vol. 26. no. 5.]. University of California.
- Curran, P. J. and Steven, M. D. (1983). Multispectral remote sensing for the estimation of green leaf area index [and discussion]. *Philosophical Transactions of the Royal Society of London. Series A, Mathematical and Physical Sciences*, 309(1508):257–270.
- De Jong, S. (1990). Multivariate calibration, h. martens and t. naes, wiley, new york, 1989. isbn 0 471 90979 3 no. of pages: 504. *Journal of Chemometrics*, 4(6):441–441.
- Denniss, A. (1995). T. m. lillesand, & r. w. kiefer, 1994. remote sensing and image interpretation, 3rd ed. xvi + 750 pp. new york, chichester, brisbane, toronto, singapore: John wiley & sons. *Geological Magazine*, 132:248–249.
- Devyver, P. and Kittler, J. (1982). *Pattern Recognition: A Statistical Approach*. Prentice-Hall.

- Ferraro, D. (2008). Data mining using k-means clustering and classification and regression trees (cart) as post-processing methods: identifying management and environmental factors for explaining sugarcane yield in northern argentina (1971-2005). *Proceedings of the iEMSs Fourth Biennial Meeting: International Congress on Environmental Modelling and Software (iEMSs 2008)*, ISBN: 978-84-7653-074-0:1959–1960.
- Galton, F. (1892). *Finger prints*. Macmillan and Company.
- Geladi, P. and Kowalski, B. R. (1986). Partial least-squares regression: a tutorial. *Analytica Chimica Acta*, 185(0):1 – 17.
- Gonzalez, R. and Woods, R. (1996). *Tratamiento digital de imágenes*. Editorial Díaz de Santos, S.A.
- Gutiérrez, C.P. y Nieto, L. (2006). *Teledetección: nociones y aplicaciones*. Carlos Pérez Gutiérrez, Ángel Luis Muñoz Nieto.
- Hartigan, J. A. (1975). *Clustering Algorithms*. John Wiley & Sons, Inc., New York, NY, USA, 99th edition.
- Hastie, T., Tibshirani, R., and Friedman, J. (2009). *The elements of statistical learning: data mining, inference and prediction*. Springer, 2 edition.
- Hebb, D. O. (1949). *The Organization of Behavior: A Neuropsychological Theory*. Wiley, New York, new ed edition.
- Hough, H. (1991). *Satellite Surveillance*. Loompanics Unlimited.
- Kohonen, T., editor (1997). *Self-organizing Maps*. Springer-Verlag New York, Inc., Secaucus, NJ, USA.
- Lehmann, E. and Casella, G. (1998). *Theory of Point Estimation*. Springer Texts in Statistics. Springer.
- Mitchell, T. M. (1997). *Machine Learning*. McGraw-Hill, Inc., New York, NY, USA, 1 edition.
- Rodriguez, J. e. a. (2010). Ieee transactions on pattern analysis and machine intelligence. *Comput. Sci. Fac., Univ. of the Basque Country (UPV-EHU), San Sebastian, Spain*.
- Samuel, A. L. (1959). Some studies in machine learning using the game of checkers. *IBM Journal of Research and Development*, 3(3):210–229.
- Sandhu, H. S., R. A. Gilbert, J. M. McCray, R. P. B. E. G. P., and Montes, G. (2012). Relationships among leaf area index, visual growth ratings, and sugarcane yield. *Journal of the American Society of Sugar Cane Technologists*, 32(32):1–14.

- Sandoval, P., González, J., de Investigación de la Caña de Azúcar de Colombia, C., and Colombia. Departamento Administrativo de Ciencia, T. e. I. (2012). *Principios y aplicaciones de la percepción remota en el cultivo de la caña de azúcar en Colombia*. Centro de Investigación de la Caña de Azúcar de Colombia.
- Shlens, J. (2005). A tutorial on principal component analysis. In *Systems Neurobiology Laboratory, Salk Institute for Biological Studies*.
- Team, R. D. C. (2014). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.
- Tenenhaus, M., Vinzi, V. E., Chatelin, Y.-M., and Lauro, C. (2005). Pls path modeling. *Computational Statistics & Data Analysis*, 48(1):159 – 205. Partial Least Square.
- W.D.Clayton, R.Govaerts, K. H. . M. (1753). *Saccharum officinarum* L., sp. pl.: 54. World Checklist of Arecaceae. Facilitated by the Royal Botanic Gardens, Kew.