Evaluación de modelos de aprendizaje estadístico aplicados en estimación de variables biofísicas en cultivos de caña de azúcar

Proyecto de Maestría

Camilo Alberto Herrera Rozo

Director
Pablo A. Cipriotti
Ingeniero Agrónomo - Dr. Ciencias Agropecuarias

Universidad de Buenos Aires
Facultad de Agronomía
Maestría en Biometría y Mejoramiento
Buenos Aires - Argentina, Junio
2014

1. Introducción

La Agricultura de precisión y 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 lo son Área foliar (AF) o Índices de área foliar (IAF) en cultivos de caña de azúcar se realiza en la actualidad por medio de procesos destructivos. El conteo de la cantidad de tallos (NT) en cultivos de caña de azúcar es tambien un proceso de estimación empírico y poco preciso; estas dos características plantean un panorama interesante de investigación con la finalidad de utilizar nuevos métodos de estimación indirecta y precisa, a los cuales la estadística pueden ser una ayuda invaluable.

En tal sentido, existen hoy una gran cantidad de métodos estadísticos que permiten realizar infinidad de análisis distintos y cubrir cualquier variable y aplicación. El uso de la inteligencia artificial (visión computacional, sistemas expertos, sistemas de ayuda de decisión, etc.) y otras técnicas potenciales (redes neuronales, lógica difusa y bioinformática) pueden proporcionar soluciones a los problemas analíticos en sistemas agrícolas complejos de manera eficaz (Bustos M, 2005). En este contexto, para analizar estructuras de datos como los que se mencionan en el párrafo anterior, tres métodos son de particular interés. Se desea 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 (Breiman, 2001) y el Aprendizaje por cuantificación vectorial (LVQ) (Kohonen, 1997). Estos tres métodos serán aplicados a imágenes multiespectrales (Hough, 1991) y datos obtenidos mediante procesos de simulación del comportamiento de cultivos de caña de azúcar.

En este proyecto de tesis se pretende integrar métodos estadísticos y sistemas de visión artificial o visión por computadora (Gonzalez and Woods, 1996), para mejorar la comprensión de procesos naturales que ocurren en los cultivos. Dichos procesos son captados por sensores remotos, y revelan patrones espaciales presentes en los cultivos (Gutiérrez, 2006). 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)], [La Regresión por mínimos cuadrados Parciales] y [Random Forest] aplicados en imágenes multiespectrales de cultivos de caña de azúcar.
- Estimar el área foliar (AF) y el índice de área foliar (IAF) en cultivos de caña de azúcar y la cantidad de tallos (NT) en cultivos de caña de azúcar.
- Ajustar modelos estadísticos y de inteligencia artificial que relacionen correctamente la información de las imágenes multiespectrales a las variables de cultivo (AF), (IAF) y (NT).

2. Revisión de la literatura

La caña de azúcar (*Saccharum officinarum*) 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. Existe una infinidad de métodos relacionados con la medición del desarrollo de los cultivos entre estos métodos encontramos nuevas aplicaciones como lo son la percepción remota o teledetección, la cual se define como el grupo de técnicas para la obtención de información confiable 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 en 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 área ocupada por hojas verdes en relación con la unidad de superficie de suelo. La información precisa y oportuna sobre IAF tiene gran importancia y aplicaciones en la agricultura para la estimación del rendimiento y la evaluación del estrés en los cultivos, y en la ecología para el estudio de la producción primaria y el cambio ambiental (Curran and 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 and 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 las máquinas de aprendizaje o el aprendizaje automático como el campo de estudio que da a las computadoras la capacidad de aprender sin estar programado explícitamente, lo que 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 arboles aleatorios y el aprendizaje por cuantificación vectorial (LVQ) son tres metodologías de mucho interés para el desarrollo de esta investigación ya que aunque son métodos cada uno desarrollado para diferentes aplicaciones en un principio son altamente escalables y se espera obtener resultados interesantes con la aplicación de métodos que con anterioridad no se han probado sobre cultivos de caña.

2.1. PLS o Regresión por mínimos cuadrados parciales

La Regresión por mínimos cuadrados 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 and Kowalski (1986) , De Jong (1990)).

Es una particular forma de análisis multivariante, este método relacionado con la regresión de componentes principales, PCR ("Principal component regression") 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 le 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. Random Forest o Árboles Aleatorios

En una serie de documentos e informes técnicos, Breiman (Breiman (1996), Breiman (2000), Breiman (2001), Breiman (2004)) demostró que hay una ganancia sustancial en la precisión en clasificación y regresión mediante el uso de conjuntos de arboles donde cada arbol en el conjunto se cultiva de acuerdo a un parametro 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 inyección de aleatoriedad, estos procedimientos son llamados "Random Forest".

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 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 ([winner-take-all] el ganador toma todo mediante el aprendizaje de Hebb) . 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).

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, aprendizaje 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

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

2.5. Validación cruzada K-veces ("K-Fold")

Sería ideal que, en caso que tuviéramos datos suficientes, dejáramos 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 repetimos el procedimiento en K ocasiones, cada una con datos distintos para entrenamiento y validación, ganamos 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:

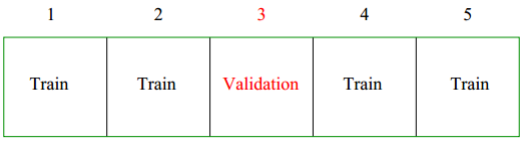

Figura 1: k-fold

Para el grafico anterior en k=3, ajustamos 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 $\varkappa : \{1,. \ . \ . \ , N\} \rightarrow \{1,. \ . \ . \ , K\}$ una función de indexación la cual que indica a que partición es asignada la observación i por la aleatorización. Denotemos por $\hat{f}^{-k}\left(X\right)$

la función de ajuste, computada con la k-esima parte de los datos eliminada, entonces la validación cruzada estimada del error de predicción es:

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

La típica elección es k = 5 o 10, pero incluso podríamos 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 de opta por selección de un K un poco más conservador como lo es 5 o 10 lo que es suficiente para obtener buenos resultados (Hastie et al., 2009).

3. Antecedentes

Hace más de 50 años que se toman fotografías a color e infrarrojo para seguir el crecimiento de las 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 rápido 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 de 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, 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, donde 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 presentado comprende varias etapas que finalmente e logra una síntesis de toda la información tanto de la imagen como de la información histórica permitiendo la normalización de todas las variables en conjunto, 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 en el 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 Lansat TM y ETM+ con las cuales generó modelos de predicción de rendimiento de caña aplicando algoritmos de random forest optimizados, logrando una estimación bastante buena para algunas de las variedades estudiadas.

4. Metodología

En el desarrollo de este proyecto de investigación se cuenta con 2 escenarios a desarrollar sobre este cultivo, donde una parte del estudio se realizara en cultivos experimentales y la otra parte del estudio se realizará sobre cultivos de caña comerciales, cada uno de estos cultivos se encuentran ubicados en el valle geográfico del río Cauca.

Los cultivos experimentales se utilizarán inicialmente por la necesidad de realizar pruebas sobre cultivos donde se puedan controlar las distintas fuentes de variación como lo son el 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 que en un cultivo comercial no es posible. 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 que han sido designados para este fin. Los cultivos comerciales serán un objetivo estratégico a evaluar, pues al depender de las políticas de plantación de los ingenios, no es posible controlar muchos de los factores relacionados con el crecimiento del cultivo.

Estos dos escenarios involucrados 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 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 experimentales, 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. Posteriormente, los cultivos comerciales servirán para realizar la validación del método desarrollado inicialmente con los cultivos experimentales.

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.

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

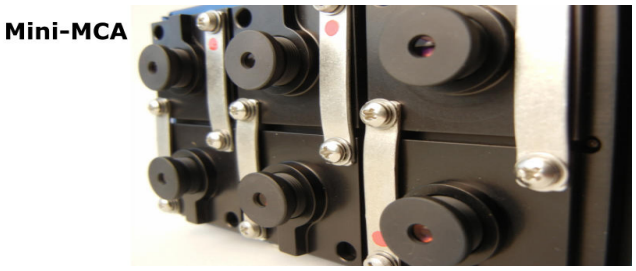

Figura 2: Cámara Multi-espectral

Se estudiará la información reflectada 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 x 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 6 bandas espectrales las cuales corresponden a un rango del espectro.

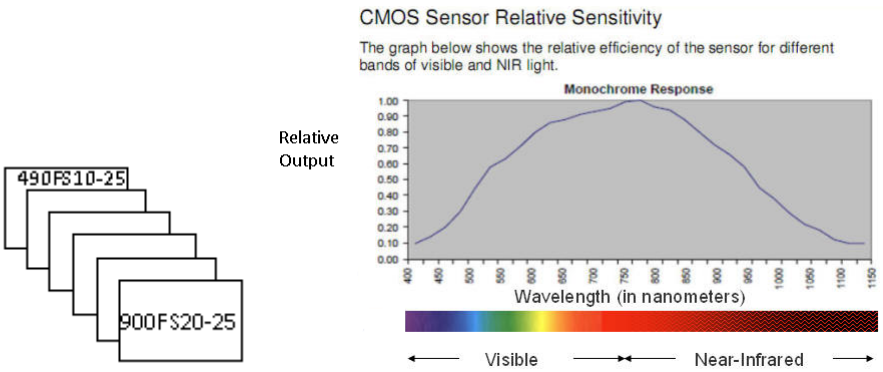

Figura 3: Rango Espectral - mini-MCA

Para el caso específico de la cámara tetracam mini-MCA de 6 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 realizara una muestra (destructiva) para el caso de la medición de las variables biofísicas. Así, con esta información de muestreo será posible realizar un proceso de validación cruzada junto a las imágenes tomadas del campo lo que permita 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 de 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 se está cambiando del paradigma de cómo obtener la información al referente 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 un gran cantidad de sensores que pueden recolectar una cantidad inmensa de información, la cual será la que alimentara nuestros modelos que finalmente realizará 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.

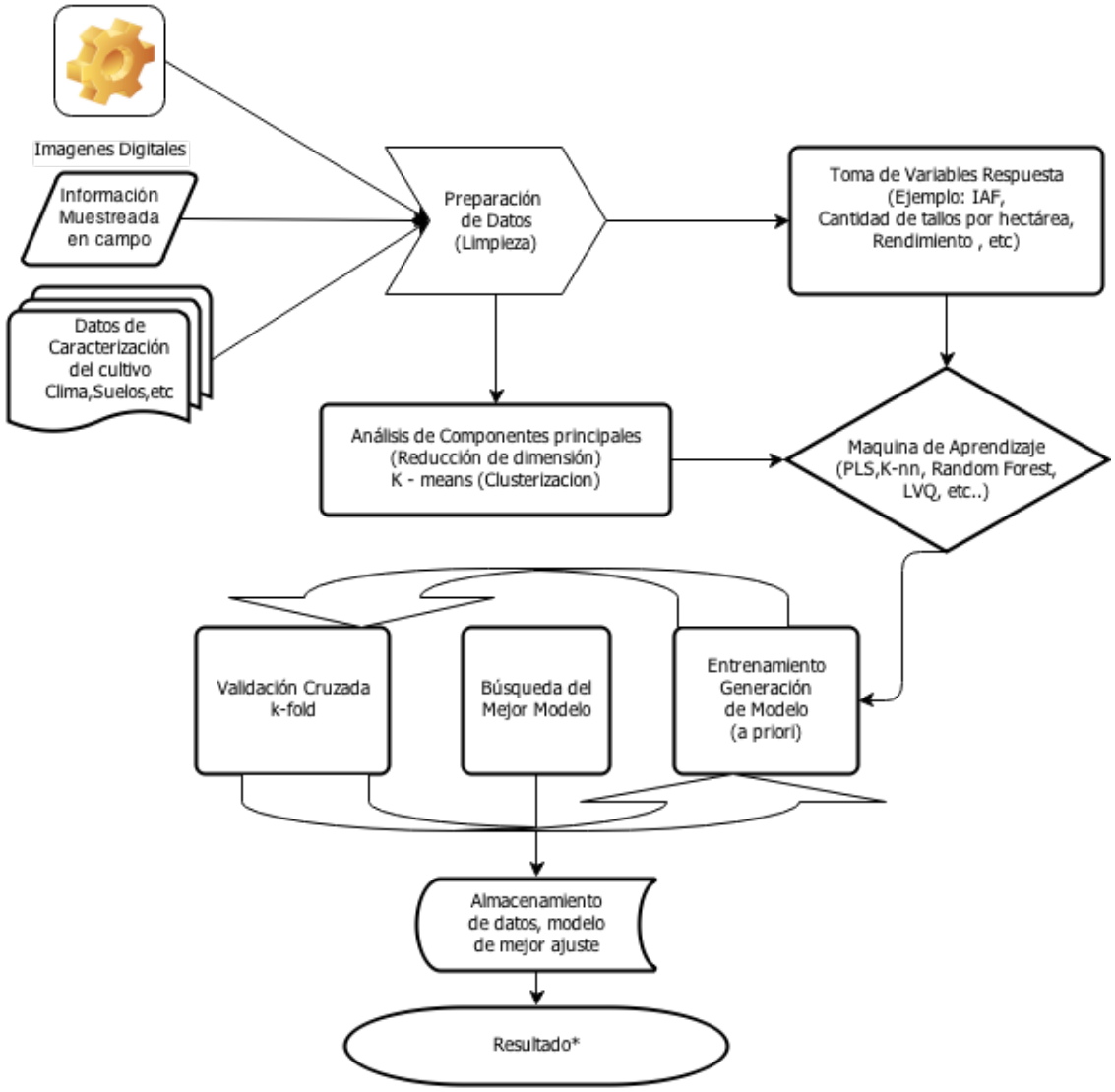

Figura 4: Propuesta Analítica Básica

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 realizaran los siguientes procesos sobre la información disponible:

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

1. Variables Iniciales

    a) Usar los valores de las bandas espectrales de las imágenes como variables.

    b) Generar índices espectrales a partir de las bandas espectrales para agregarlos como variables.

    c) 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.

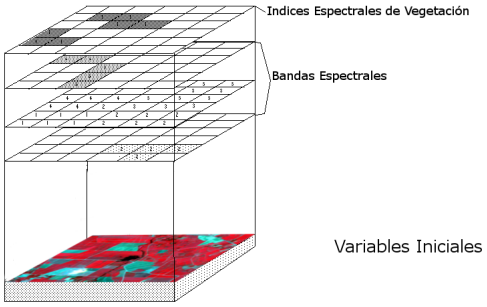

Figura 5: Variables Iniciales

2. Variables Respuesta

    a) Recopilar información de Área Foliar e IAF por muestreo

    b) Recopilar información de cantidad de tallos por lote productivo

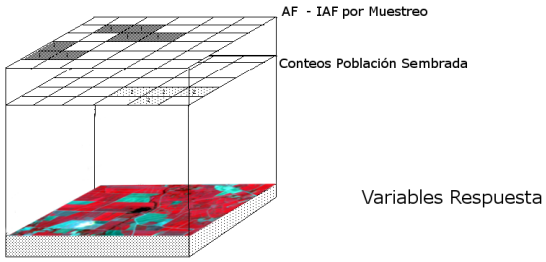

Figura 6: Variables Respuesta

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

1. Estandarizar toda la información al tamaño de pixel adecuado para maximizar la eficiencia del proceso y manejar la misma resolución espacial en todo el proceso.
2. Se realizará un proceso de reducción de dimensionalidad considerando distintos sets de variables. [Mediante Análisis de Componentes Principales (Shlens, 2005)]
3. Realizar un proceso de clusterización o agrupamiento de pixeles con características de respuesta espectral ajustando pixeles 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:

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

Después de realizar estos procesos previos existen 2 posibilidades para abordar el problema:

1. Usar solo los pixeles 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.
2. Realizar una interpolación en las variables respuesta a pixeles 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 o PLS , Los Arboles Aleatorios o Random Forest y el Aprendizaje por cuantificación vectorial (LVQ) son metodologías de un gran interes 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 variables a estimar AF , IAF y NT, y se validara mediante un proceso de validación cruzada.

Una vez se han evaluado todas estas alternativas, se seleccionará el método que permita realizar una mejor estimación de las variables biofísicas y de cantidad 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 las imágenes multiespectrales en términos de la respuesta observada. 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 datos de prueba. Se utiliza en entornos donde el objetivo principal es la predicción y se quiere estimar cómo de preciso es un modelo que se llevará a cabo a la práctica.(Devyver and Kittler, 1982) Es una técnica muy utilizada en proyectos de inteligencia artificial para validar modelos generados.

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 muestreable tanto mediante toma de imágenes como de muestreadores en campo, además de considerar la capacidad de análisis de la información muestreada. Por último, se evaluará la resolución final mediante la cual se realizará el análisis de los datos, esto se definiría intentando lograr un balance entre la ganancia en resolución y la capacidad computacional para realizar el procesamiento de los datos, 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 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, 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. 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 las condiciones anormales en los cultivos, de manera que se pueden tomar medidas de control o de prevención oportunas. El uso de 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 especifico 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 fumigacion, UAVs, grúas u otros y la información puede ser procesada facilmente una vez se establezca un protocolo estandar 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 mayor resolución en imágenes multiespectrales unido a sistemas de SIG es posible generar mapas mucho más detallados de importantes características, tanto del cultivo como del comportamiento 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 población como biofísicas, 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

El trabajo se desarrollará contando con información y datos a portados por el programa de agronomía del Centro de Investigación de la Caña de Azúcar de Colombia y el área de biometría de esta entidad. Los gastos adicionales distintos a la toma de información (computo, papelería , etc.) 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 (Saccharun 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.

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.

Hough, H. (1991). *Satellite Surveillance*. Loompanics Unlimited.

Kohonen, T., editor (1997). *Self-organizing Maps*. Springer-Verlag New York, Inc., Secaucus, NJ, USA.

Mitchell, T. M. (1997). *Machine Learning*. McGraw-Hill, Inc., New York, NY, USA, 1 edition.

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*.

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.