

Análisis de regresión para el cálculo de biomasa en el departamento de Nariño (Colombia) utilizando imágenes satelitales Landsat

GIEE
Universidad de Nariño
San Juan de Pasto, Colombia
Email:

GIEE
Universidad de Nariño
San Juan de Pasto, Colombia
Email:

Abstract—En este artículo se describe la metodología utilizada para la construcción de un primer mapa del potencial de biomasa en el departamento de Nariño (Colombia) a partir de imágenes satelitales Landsat de libre acceso. Se analizan las diferentes bandas de las imágenes satelitales disponibles y su relación con bases de datos previas de biomasa aplicando diferentes técnicas de regresión y obteniendo un modelo para la generación de mapas actualizados en el departamento de Nariño.

Keywords—biomass, regression models

I. INTRODUCCIÓN

En las últimas décadas la investigación en fuentes alternativas de energía ha recibido particular atención y ha pasado de tener un alto interés en los círculos académicos a convertirse en un punto prioritario en la agenda de gobiernos y organizaciones a nivel mundial. La dependencia en combustibles fósiles y acuerdos internacionales como el protocolo de Kyoto han impulsado aún más el interés alrededor del tema. En particular, la implementación de soluciones como paneles fotovoltaicos, parques eólicos y plantas de biomasa han atraído gran atención en especial en zonas con baja o nula cobertura.

En Colombia, según estudios del Ministerio de Minas y Energía, en el departamento de Nariño hay 15 municipios con cobertura eléctrica inferior al 80% [9]. Como nueva estrategia para enfrentar esta problemática se ha propuesto el estudio y análisis de fuentes alternativas de energía en la zona. Uno de los principales objetivos es la medición y estimación de potenciales energéticos para identificar las zonas más viables en la región donde efectuar pruebas piloto y estudios de factibilidad.

Sin embargo, uno de los principales retos para la ubicación de dichas zonas es la ausencia de bases de datos actualizadas así como series de tiempo históricas que apoyen el proceso de toma de decisiones. Igualmente, restricciones de tiempo y costos impiden el despliegue de trabajo de campo para la recolección de información. En el caso del análisis del potencial biomásico, estas restricciones se acentúan debido a la toma manual de muestras, extensión del área de estudio, análisis de laboratorio, dificultad del terreno e, incluso, presencia de grupos armados en la zona.

En este sentido, diversas investigaciones han demostrado la utilidad del uso de imágenes satelitales para la generación de modelos que permitan calcular la cantidad de biomasa presente

en un determinado lugar. Desde hace más de 30 años, se cuenta con acceso al repositorio de imágenes satelitales Landsat [17] de manera libre y gratuita. Bajo el debido tratamiento, estas imágenes pueden ser usadas para calcular valores nominales de biomasa a partir de modelos de regresión y trabajo de campo. Sin embargo, dadas las dificultades para realizar dicho trabajo de campo, este estudio propone utilizar imágenes provistas por investigaciones anteriores. [3] y [2] proporcionan bases de datos de los índices de biomasa a nivel pan-tropical entre los años 2000 y 2003. Al igual que el conjunto de imágenes Landsat, las imágenes georreferenciadas para cada uno de los países analizados son de libre acceso y se encuentran disponibles en [5].

Esta investigación presenta la metodología propuesta para la generación de un modelo de predicción de biomasa, basado en modelos de regresión e imágenes satelitales de libre acceso, y su extrapolación al resto del área de estudio.

El área de estudio de esta investigación fue el departamento de Nariño, el cual está ubicado en el extremo Suroccidental de Colombia (en la frontera con Ecuador) con una extensión aproximada de 33.268 km, una población de 1,702 millones (según el censo de 2013) y ubicada entre coordenadas 00° 31' 08" y 02° 41' 08" Norte y 76° 51' 19" y 79° 01' 34" Oeste (figura 1).

II. TRABAJOS RELACIONADOS

El estudio de índices de biomasa ha sido ampliamente registrado por diferentes estudios. Muchos de ellos demuestran la utilidad del uso de imágenes satelitales a diferentes resoluciones. Por lo general, estos estudios parten de un trabajo de campo donde se calcula el valor nominal de biomasa a diferentes muestras tomadas de manera manual y haciendo uso de técnicas tradicionales de laboratorio. Posteriormente, se utilizan estos resultados y las diferentes bandas proveídas por las imágenes de satélite para inferir un modelo utilizando alguna técnica de regresión que después es extrapolada al resto del área de estudio.

Por ejemplo, [12] usa esta metodología para detectar cambios en los niveles de biomasa en diferentes zonas costeras de los Estados Unidos utilizando imágenes LIDAR¹ y regresión

¹Una técnica de muestreo que mide distancia haciendo uso de rayos láser.

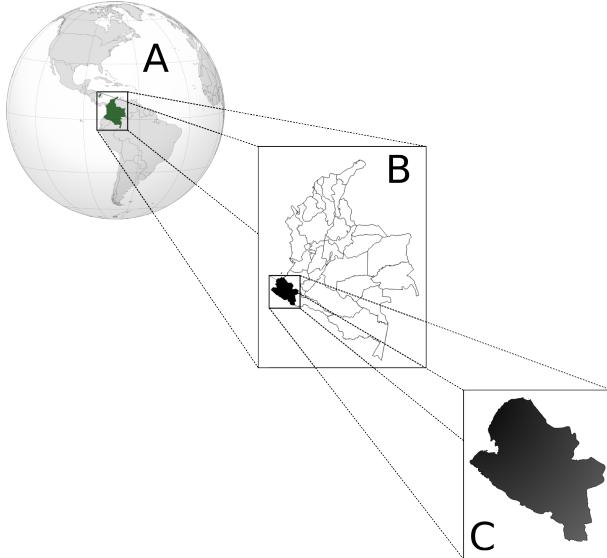


Fig. 1. Localización área de estudio

lineal. De forma análoga, [3] usa imágenes MODIS² y árboles de decisión (en adición a las técnicas tradicionales de regresión) para estimar el índice AGB (Above-Ground Biomass) en una extensa área del África tropical. Similar a este trabajo, [14] utilizan imágenes de radar para predecir AGB en cuatro reservas y parques nacionales africanos clasificando diferentes tipos de corteza terrestre.

[15] hacen también uso de imágenes MODIS en conjunto con imágenes ASTER³ para estimar biomasa con el fin de levantar un inventario de captura de carbono. Un aporte importante de esta publicación es que comparten la metodología utilizada durante el proyecto. [16] introducen el uso de nuevas técnicas de regresión (reduced major axis regression, gradient nearest neighbor imputation y ramdom forest regression trees) para la generación de modelos de biomasa esta vez analizando imágenes Landsat.

En [10] se introduce bioSTRUCT, un método para generar correlaciones entre los valores continuos medidos por las bandas de las imágenes satelitales y el AGB medido previamente usando técnicas de laboratorio. El artículo ilustra la metodología con un caso de estudio en Alberta (Canadá) e imágenes Landsat ETM+ de libre acceso. Como resultado se obtienen formulas de regresión a partir de un número limitado de muestras que pueden extrapolarse al resto del área de estudio.

III. METODOLOGÍA

El repositorio de imágenes satelitales Landsat es amplio y diverso. Si bien se constituye como una gran herramienta para la comunidad científica, su uso requiere un tratamiento previo. De igual manera, la construcción de un modelo preliminar de biomasa a partir de información secundaria exige la selección y validación de diferentes técnicas de regresión disponibles.

²Moderate Resolution Imaging Spectroradiometer - <http://modis.gsfc.nasa.gov/>

³Advanced Spaceborne Thermal Emission and Reflection Radiometer - <http://asterweb.jpl.nasa.gov/>

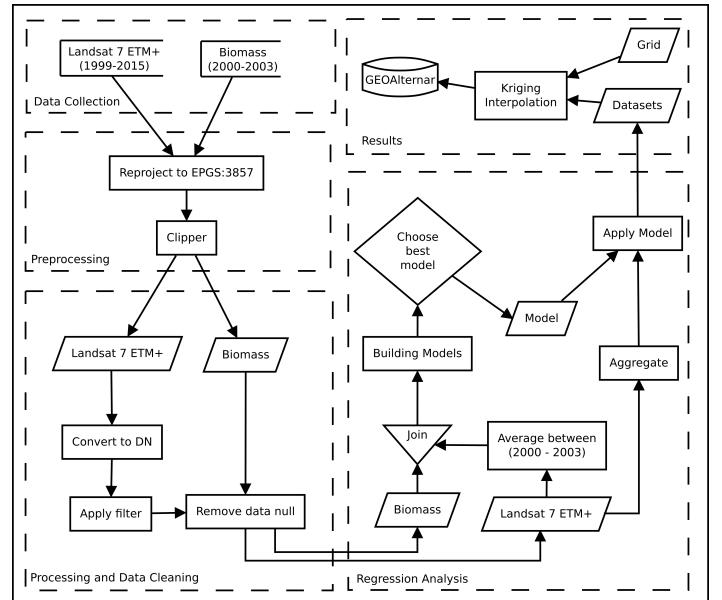


Fig. 2. Metodología

Esta sección resume una metodología de cinco etapas para la construcción del modelo de biomasa para el departamento de Nariño. La figura 2 ilustra la metodología propuesta. A continuación se explica en más detalle cada una de las etapas.

A. Obtención de datos

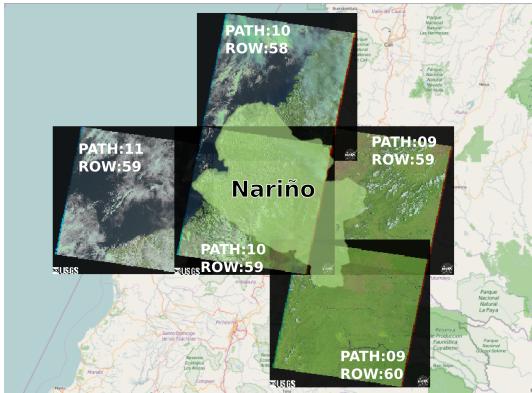
El proceso de obtención de datos se realizó tomando imágenes satelitales proveidas por el sensor Landsat 7 ETM+. En este proceso se descargaron 1362 imágenes satelitales desde el año 1999 hasta mediados del año 2015. Para cubrir el departamento en su totalidad fue necesario descargar imágenes de cinco escenas diferentes. La figura 3a detalló los respectivos identificadores (Path ID y Row ID) y extensión de cada una de las escenas usadas.

Igualmente, durante esta etapa se tuvo acceso al mapa de biomasa construido por [3]. Este es un mapa con resolución espacial de 1 Km^2 construido a partir de un modelo basado en imágenes MODIS recolectadas durante el año 2000 y 2003.

B. Preprocesamiento

En esta etapa se realizó un trabajo básico de procesamiento sobre las imágenes adquiridas. Primero, dada la extensión del área de estudio, las escenas descargadas tenían diferentes sistemas de coordenadas (EPSG:32618 y EPSG:32617). Por motivos de visualización se decidió unificar el sistema de coordenadas usando EPSG:3857, popular entre las herramientas de mapeo y desarrollo de aplicaciones web. Muchas de las escenas cubrían una gran área del Océano Pacífico así como de otros departamentos de la región. Se recortó las imágenes para contener solo los datos referentes al departamento de Nariño. La figura 3b ilustra el resultado final de esta etapa.

Un proceso similar se realizó para el mapa de biomasa, como se muestra en la figura 4.



(a) Imágenes Satelitales de Nariño



(b) Imágenes recortadas de Nariño

Fig. 3. Preprocesamiento

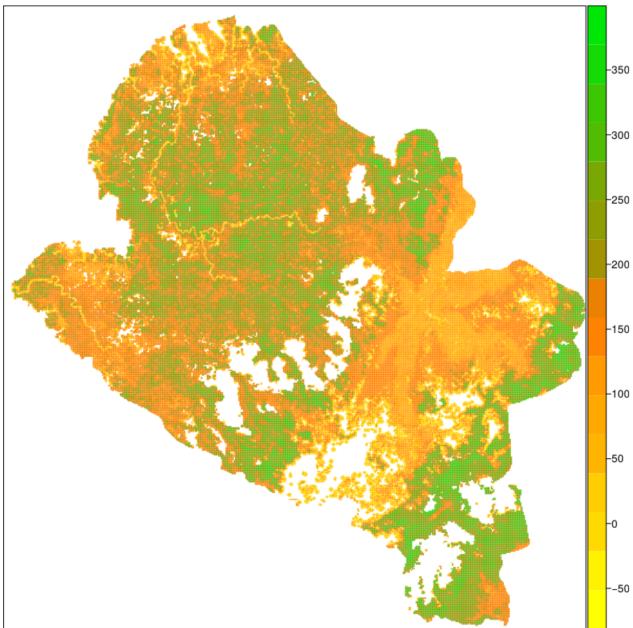


Fig. 4. Mapa de biomasa en Nariño de 2000-2003 [3]. Los valores se expresan en toneladas por hectárea ($Mgha^{-1}$).

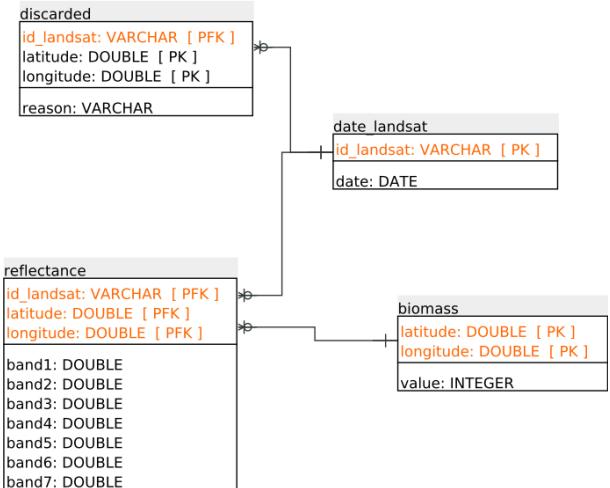


Fig. 5. Modelo entidad-relación Landsat

C. Procesamiento y limpieza de datos

Con el objetivo de organizar los datos adquiridos se diseñó una base de datos a partir de cuatro entidades fundamentales: la fecha de la toma, los valores de reflectancia provenientes de cada imagen, los valores correspondientes de biomasa y las ubicaciones descartadas durante la limpieza de datos. La figura 5 ilustra el diseño de la base de datos y los detalles de cada tabla se comentan a continuación:

- Tabla date_landsat: en la cual se almacenan las fechas de las imágenes satelitales.
- Tabla reflectance: en la cual se almacenan los datos capturados y convertidos a reflectancia, de las bandas Landsat (1-5 y 7) y la temperatura en grados kelvin de la banda 6.
- Tabla discarded: en la cual se almacenan los datos que fueron descartados por diversas razones (son puntos nublados, datos ambiguos o no corresponden a vegetación).
- Tabla biomass: en la cual se almacenan los datos de biomasa extraídos de [3].

El procesamiento de las imágenes y alimentación de la base de datos se realizó a través de scripts y archivos procesados por lotes. Entre los procesos realizados se transformó los valores originales extraídos de las imágenes (o digital numbers) a su correspondiente valor de reflectancia. Se recomienda usar valores de reflectancia dado que esta magnitud, al ser el radio entre la luz emitida y reflejada, resulta menos susceptible a los efectos de sombras o inclinación solar [6].

Se utilizó el algoritmo propuesto en [11] para detectar puntos nublados en la zona clasificándolos como nubes frías, calientes o ambiguas. Estos puntos se almacenaron con la intención de realizar posteriores estudios de nubosidad. Finalmente, se aplicó un filtro adicional para calcular el índice NDVI (Normalized Difference Vegetation Index) con el objetivo de trabajar solo con aquellos puntos relacionados con vegetación y excluir áreas como cuerpos de agua o ciudades. La tabla I muestra la relación de los datos obtenidos en este proceso.

TABLE I. DATOS OBTENIDOS EN EL PROCESAMIENTO Y LIMPIEZA DE DATOS

| Nombre | Valor | Detalle |
|-----------------------------|------------|--|
| Datos biomasa | 81.993 | Registros biomasa (años 2000 a 2003 [3]) |
| Datos biomasa usados | 140018 | construcción del modelo |
| Imágenes Landsat procesadas | 1321 | Imágenes de Nariño (2008 a 2014) |
| Nube caliente | 3.731.768 | Registros 2000 a 2014 |
| Nube Fria | 27.827.009 | Registros 2000 a 2014 |
| No vegetación | 3.459.210 | Registros 2000 a 2014 |
| Ambiguo | 11.987.340 | Registros 2000 a 2014 |
| Total Datos Descartados | 47.005.327 | Total datos descartados |
| Datos Validos Reflectance | 4.071.185 | Registros 2000 a 2014 |
| Datos Totales | 51.076.512 | Registros Totales (año 2000 a 2014) |

D. Análisis de regresión

El análisis de regresión se realizó tomando los valores de las bandas Landsat obtenidas entre los años 2000 y 2003 y adicionando los valores correspondientes de biomasa para cada ubicación. Dichos valores se extrajeron de [3]. Para poder obtener una mayor confiabilidad en el modelo solo se tuvo en cuenta aquellas ubicaciones con un número significativo de muestras. Dada la alta nubosidad de la zona, muchos de los puntos contaban con pocas lecturas. Por lo tanto, se consolidó un nuevo conjunto de datos con el promedio de aquellos puntos que superaban al menos un número considerable de lecturas válidas.

Se construyeron diferentes modelos, iterando el número de muestras por cada punto entre 10 y 45. El mejor modelo se obtuvo cuando el número de muestras superaba las 35 lecturas. El conjunto de datos final arrojó 1009 registros válidos. El comportamiento en las demás iteraciones indica que con pocas muestras el conjunto de entrada es altamente heterogéneo, guiado por aquellos puntos con pocas lecturas y alta variabilidad. En cambio, al aumentar el número de lecturas, la variabilidad de las muestras baja pero con el alto riesgo de sobrecargar el modelo [1].

Antes de aplicar las técnicas de regresión, se procedió a evaluar la calidad y relevancia de las bandas de las imágenes Landsat con el objetivo de predecir valores de biomasa. Se utilizó el algoritmo propuesto en [13] para la extracción y evaluación de atributos. El algoritmo está diseñado como un recubrimiento alrededor del algoritmo de clasificación random forest y califica cada atributo en el conjunto de datos de acuerdo a su importancia a la hora de clasificar el atributo buscado. En la figura 6 se puede observar la relevancia de todas las bandas Landsat (en verde) que se ubican por encima de los valores por defecto (en azul). Esto indica que la relevancia de las bandas de las imágenes Landsat esta por encima del azar. Concluimos que todas las bandas resultan importantes a la hora de modelar biomasa.

Con un nuevo conjunto de datos definido y evaluado, se continuó con la construcción de modelos de regresión utilizando diversas técnicas de análisis. La biblioteca de código abierto rminer, presentada por [7] para la herramienta R, provee diferentes implementaciones y una interfaz que facilita la ejecución de diferentes pruebas y la extracción de modelos y sus correspondientes métricas de evaluación.

Se construyó un total de 13 modelos y se evaluaron 6 métricas por cada uno. La tabla II ilustra los resultados

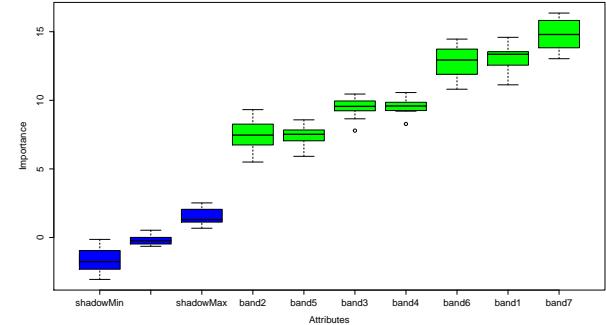


Fig. 6. Relevancia de bandas Landsat en el análisis de regresión

TABLE II. MÉTRICAS DE MODELOS ANALIZADOS. LOS VALORES EN NEGRILLA INDICAN LOS MEJORES RESULTADOS PARA CADA MÉTRICA.

| | SAE | MAE | RAE | RMSE | COR | R2 |
|--------------|--------------------|-----------------|-----------------|-----------------|----------------|----------------|
| ctree | 10406.58225 | 30.88007 | 65.04650 | 40.02893 | 0.69401 | 0.48165 |
| rpart | 10197.95826 | 30.26100 | 63.74249 | 39.37592 | 0.70520 | 0.49730 |
| kknn | 9147.51425 | 27.14396 | 57.17667 | 36.86581 | 0.74955 | 0.56182 |
| mlp | 9179.79310 | 27.23974 | 57.37843 | 34.70711 | 0.78122 | 0.61031 |
| mlpe | 8746.27740 | 25.95335 | 54.66874 | 34.57953 | 0.78309 | 0.61223 |
| ksvm | 8462.61487 | 25.11162 | 52.89570 | 34.67742 | 0.79830 | 0.63729 |
| randomForest | 8807.76477 | 26.13580 | 55.05306 | 34.70615 | 0.78239 | 0.61214 |
| mr | 10410.13919 | 30.89062 | 65.06873 | 38.61068 | 0.72000 | 0.51840 |
| mars | 8842.91866 | 26.24011 | 55.27279 | 33.96852 | 0.79161 | 0.62665 |
| cubist | 9012.54150 | 26.74345 | 56.33302 | 35.70576 | 0.77611 | 0.60235 |
| pcr | 10337.63121 | 30.67546 | 64.61552 | 38.59290 | 0.72023 | 0.51873 |
| plsr | 10337.63121 | 30.67546 | 64.61552 | 38.59290 | 0.72023 | 0.51873 |
| cppls | 10337.63121 | 30.67546 | 64.61552 | 38.59290 | 0.72023 | 0.51873 |

obtenidos. Las técnicas utilizadas fueron: ctree (conditional inference tree), rpart (decision tree), kknn (k-nearest neighbor), mlp (multilayer perceptron with one hidden layer), mlpe (multilayer perceptron ensemble), ksvm (support vector machine), randomForest (random forest algorithm), mr (multiple regression), mars (multivariate adaptive regression splines), cubist (rule-based model), pcr (principal component regression), plsr (partial least squares regression) y cppls (canonical powered partial least squares). Las métricas evaluadas fueron: SAE (sum absolute error), MAE (mean absolute error), RAE (relative absolute error), RMSE (root mean squared error), COR (correlation) y R2 (coefficient of determination R^2).

A partir de la tabla II y analizando los resultados de las diferentes métricas, las evaluaciones obtenidas al aplicar la técnica de support vector machines (ksvm) resultan ser las más confiables. Este modelo será utilizado en la posterior etapa para la construcción y visualización de mapas.

E. Construcción de mapas

Para la construcción de mapas de biomasa se utilizó el método Kriging ([4], [8]). Kriging provee una solución al problema de la estimación basada en un modelo continuo de variación espacial estocástica. El objetivo de Kriging es el de estimar el valor de una variable aleatoria, en este caso biomasa, en uno o más puntos no muestreados o sobre grandes bloques.

El método Kriging recibe como entrada un muestreo de datos y una malla dependiendo de la resolución que se quiera obtener. Para ello, se obtuvo una muestra de los datos obtenidos al aplicar el modelo seleccionado a datos agregados por mes y año y un mapa general que abarca el periodo de estudio. La malla se construyó con puntos regulares espaciados cada 450 metros. En las figuras 7, 8 y 9 se muestran los mapas

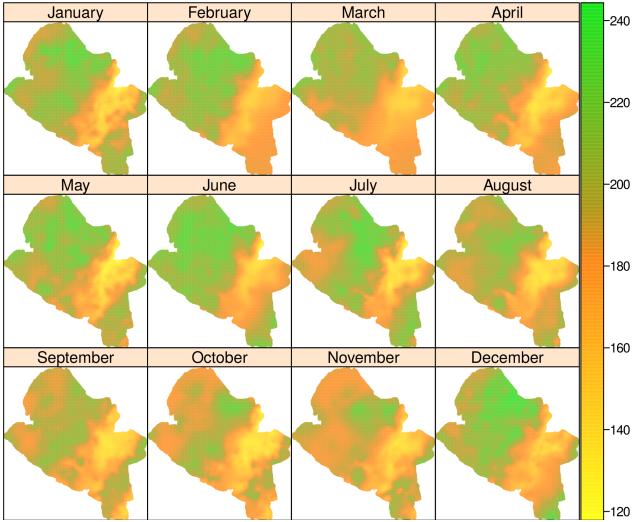


Fig. 7. Mapas de biomasa por mes

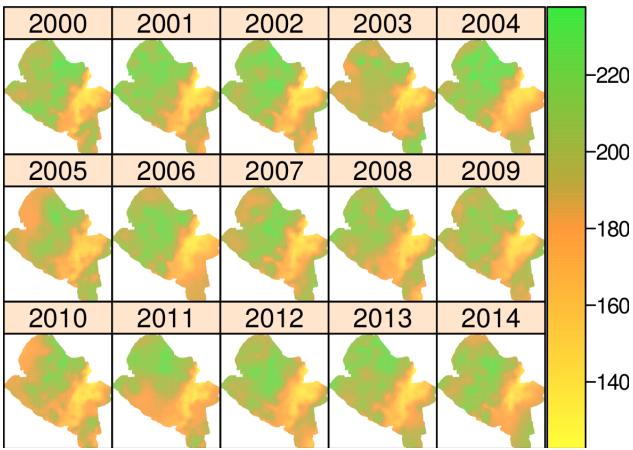


Fig. 8. Mapas de biomasa por año

obtenidos por meses, años y general respectivamente. Los valores de biomasa están expresados en toneladas por hectárea ($Mgha^{-1}$).

AGRADECIMIENTOS

Esta investigación se hizo posible gracias a los recursos otorgados por el Sistemas General de Regalías en el marco de proyecto “Análisis de Oportunidades Energéticas con Fuentes Alternativas en el Departamento de Nariño” ejecutado por el programa de Ingeniería Electrónica de la Universidad de Nariño.

REFERENCIAS

- [1] Michael A. Babyak. What you see may not be what you get: a brief, nontechnical introduction to overfitting in regression-type models. *Psychosomatic medicine*, 66(3):411–421, 2004.
- [2] A. Baccini, S. J. Goetz, W. S. Walker, N. T. Laporte, M. Sun, D. Sulla-Menashe, J. Hackler, P. S. A. Beck, R. Dubayah, M. A. Friedl, S. Samanta, and R. A. Houghton. Estimated carbon dioxide emissions from tropical deforestation improved by carbon-density maps. *Nature Climate Change*, 2(3):182–185, January 2012.
- [3] A. Baccini, N. Laporte, S. J. Goetz, M. Sun, and H. Dong. A first map of tropical africa’s above-ground biomass derived from satellite imagery. *Environmental Research Letters*, 3(4):045011, oct 2008.
- [4] Roger S. Bivand, Edzer Pebesma, and Virgilio Gómez-Rubio. *Applied Spatial Data Analysis with R*. Springer New York, New York, NY, 2013.
- [5] Woods Hole Research Center. Pantropical national level carbon stock dataset. http://www.whrc.org/mapping/pantropical/carbon_dataset.html. Accessed: 2015-08-15.
- [6] Gyanesh Chander, Brian L. Markham, and Dennis L. Helder. Summary of current radiometric calibration coefficients for Landsat MSS, TM, ETM+, and EO-1 ALI sensors. *Remote sensing of environment*, 113(5):893–903, 2009.
- [7] Paulo Cortez. Data mining with neural networks and support vector machines using the rminer tool. In *Advances in data mining. Applications and theoretical aspects*, pages 572–583. Springer, 2010.
- [8] Noel Cressie. *Statistics for Spatial Data*. Wiley-Interscience, Hoboken, NJ, 2 edition edition, July 2015.
- [9] Ministerio de Minas y Energia. Plan Indicativo de Expansión de la Cobertura del Servicio de Energía Eléctrica 2006 - 2010. Technical report, Unidad de Planeación Minero Energética - Ministerio de Minas, Colombia, 2008.
- [10] R.J. Hall, R.S. Skakun, E.J. Arsenault, and B.S. Case. Modeling forest stand structure attributes using landsat ETM+ data: Application to mapping of aboveground biomass and stand volume. *Forest Ecology and Management*, 225(1-3):378–390, apr 2006.
- [11] Richard R Irish. Landsat 7 automatic cloud cover assessment. In *AeroSense 2000*, pages 348–355. International Society for Optics and Photonics, 2000.
- [12] Victor Klemas. Remote sensing of coastal wetland biomass: An overview. *Journal of Coastal Research*, 290:1016–1028, sep 2013.
- [13] Miron B Kursa, Witold R Rudnicki, et al. Feature selection with the boruta package, 2010.
- [14] E. T. A. Mitchard, S. S. Saatchi, I. H. Woodhouse, G. Nangendo, N. S. Ribeiro, M. Williams, C. M. Ryan, S. L. Lewis, T. R. Feldpausch, and P. Meir. Using satellite radar backscatter to predict above-ground woody biomass: A consistent relationship across four different african landscapes. *Geophysical Research Letters*, 36(23), 2009.
- [15] P. Muukkonen and J. Heiskanen. Biomass estimation over a large area based on standwise forest inventory data and ASTER and MODIS

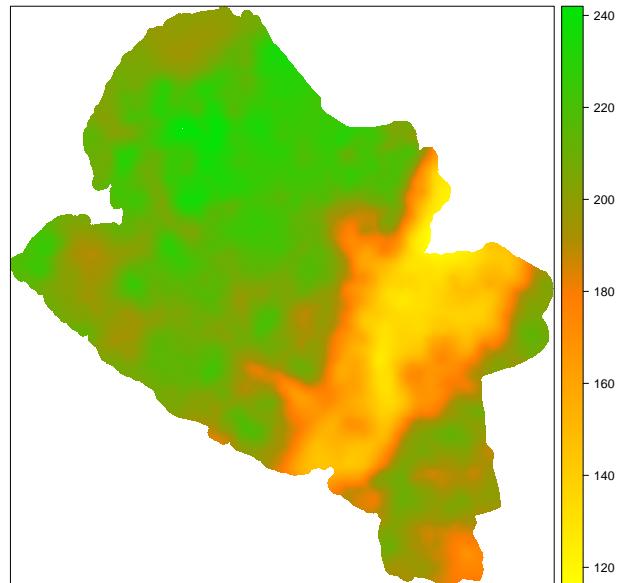


Fig. 9. Mapa de biomasa general entre los años 2000 y 2014

- satellite data: A possibility to verify carbon inventories. *Remote Sensing of Environment*, 107(4):617–624, apr 2007.
- [16] Scott L. Powell, Warren B. Cohen, Sean P. Healey, Robert E. Kennedy, Gretchen G. Moisen, Kenneth B. Pierce, and Janet L. Ohmann. Quantification of live aboveground forest biomass dynamics with landsat time-series and field inventory data: A comparison of empirical modeling approaches. *Remote Sensing of Environment*, 114(5):1053–1068, may 2010.
- [17] U.S. Geological Survey. Landsat missions. <http://landsat.usgs.gov/>. Accessed: 2015-08-15.