

АЛГОРИТЪМ, БАЗИРАН НА ЕНТРОПИЯ, ЗА РЕГИСТРИРАНЕ НА ПИКОВИ СТОЙНОСТИ ПРИ НАБЛЮДЕНИЯТА НА ИНДЕКСА НА ЛИСТНАТА ПЛОЩ (LAI) ПОСРЕДСТВОМ MODIS (NASA)

ENTROPY-BASED ALGORITHM FOR DETECTING OUTLYING OBSERVATIONS IN MODIS (NASA) LAI TIME SERIES

Г. Лечов

Аграрен Университет, Катедра „Математика, Информатика и Физика“, 4000 Пловдив, ул. „Менделеев“ 12, тел. 032/654262

letchov@gmail.com

Г. Врагов

Институт по математика и информатика, БАН, 4000 Пловдив, бул. „Руски“ 139, тел.: 032/625008

vragov@gmail.com

Abstract: Leaf area index (LAI) is a state parameter in Large-scale ecosystem models describing the exchange of fluxes of energy, mass (e.g., water and CO₂), and momentum between the surface and the planetary boundary layer. The identification and analysis of the outliers is important for the further processing of the LAI time series. Here an entropy-based algorithm was applied to identify and analysis of the outliers in LAI time series produced by MODIS (NASA) spectroradiometers over Rhodope Mountains, during 2010. The relationships between entropy and the signal quality in terms of quality control (QC) flags attach to any value of LAI in the MODIS data products are analyzed. The algorithm is implemented in C #.

Key words: leaf area index, MODIS, outliers, outages, entropy

ВЪВЕДЕНИЕ

През последните двадесет години бе отбелязан силен напредък в приложението на спътниковите дистанционни наблюдения за мониторинг на растителната покривка (РП) в локален и глобален мащаб. Това бе предизвикано от проблемите, свързани с парниковия ефект и климатичните промени.

Индексът на листната площ (Leaf Area Index - LAI) се дефинира [1] като едностранна площ на листата, намиращи се над единица земна площ. LAI е ключова променлива, която участва при формирането на потоците на обмен на енергия, маса (вода, CO₂ и импулс) в планетарния граничен слой.

През последното десетилетие сателитните наблюдения се използват за глобален мониторинг на динамиката на LAI чрез сензори със средна резолюция, като MODIS (Moderate Resolution Imaging Spectroradiometer), VEGETATION, MERIS (Medium Resolution Imaging Spectrometer) или AVHRR (Advanced Very High Resolution Radiometer). Качеството на регистрирания сигнал зависи от много фактори - наличие на облачна покривка, състояние на детекторите, сходимост на алгоритмите за идентификация на параметрите на растителната покривка, геометрията на системата слънце-сензор-земна повърхност. Това води до намаляване на амплитудата на регистрирания сигнал и получаване на занижени стойности на LAI. Редуцираното качество на сигнала, налага предварителната му обработка за отстраняване на шума във времевите серии от данни. Предварителната обработка на сигнала се извършва обикновено на два етапа. Първо се идентифицират и отстраняват пиковите (т.н. shot noise). Втората стъпка е свързана с

елиминирание на финия шум в сигнала, останал след отстраняването на пиковите стойности. За отстраняване на шума във времевите серии, получени в резултат на дистанционни наблюдения на РП, най-често се използват глобални филтри - обикновено регресионни модели (полиномиални, логистични, Гаусови и др.). Предимството им е, че с тях се отстраняват както пиковите стойности, така и фините шумове. Но ако се приложат директно, без предварителна обработка на сигнала (отстраняване на пиковите), този тип филтри може да изкривят силно реалния сигнал и да доведат до сериозни грешки в интерпретацията на наблюденията.

Заради важността на този проблем при обработката на сигнали, досега са разработени много методи за регистриране на пикове и тяхното отстраняване [1]. В научната литература обаче няма достатъчно изследвания върху пиковите във времевите серии от сателитни наблюдения на растителната покривка, като наличните са свързани предимно с измерванията на Normalized Difference Vegetation Index (NDVI). Пиковите във времевите серии данни, получени при дистанционни наблюдения на РП, са отрицателно отместени, т.е. пиковите стойности на измерваната величина (NDVI, LAI), са много по-малки от измерените. Това се дължи на наличието на облачност и промени в състава на атмосферата над наблюдаваната територия. Наред с този отрицателно отместен шум, съществува и шум в сигнала дължащ се на други фактори - силна насеченост на терена и технически проблеми със сензорите. Johnson P. and L. Eklundh[2] използват два метода за регистриране на пикови стойности при измерване на NDVI. Първият метод използва флагите за състояние на облачната покривка,

които съпровождат продуктите, включващи измерените стойности на NDVI. За отстраняване на останалите пикови стойности, те предлагат три различни метода. В първия метод, стойностите, които се различават съществено от медианата на най-близките съседи в даден прозорец се регистрират като пикове. При втория метод, сигналът се разлага на три компоненти (STL-decomposition) - сезонна компонента, тренд и шум и се прилага непараметрична регресия LOESS (Locally Weighted Scatterplot Smoothing) [3]. А третият метод използва едновременно STL-decomposition и придружаващите флагове за облачност. За отстраняване на отрицателно отместения шум в измерванията на NDVI, Xiao и др.[4] прилагат неколкостепенен филтър от тип Savitzky-Golay. В [5] Hird J. N. и G. J. McDermid правят оценка на шест различни метода за редуциране на шума в NDVI.

Пиковите стойности на LAI, все още не са добре изследвани. В [6] Kandasamy и другите автори сравняват шест метода за изглаждане на сигнала, регистриран при дистанционни наблюдения на РП. Методите са приложени към LAI, измерен посредством спектрорадиометъра MODIS за периода 2000–2008 г. над 25 места, отразяващи различни специфики на РП.

В настоящата работа е приложен непараметричен метод Palshikar (2009) [7] за регистриране на пиковите стойности на индекса на листната площ LAI, получен при глобален мониторинг на състоянието на РП, осъществяван с помощта на спектрорадиометъра MODIS на сателитите Terra и Aqua (NASA). Анализирани са наблюденията през 2010 г. над повече от 50000 хектара широколистни гори в Родопите. При широколистните гори шумът в регистрирания сигнал е най-голям. В основата на метода стои изследването на изменението на ентропията на сигнала в околностите на пиковите стойности.

До сега този метод е прилаган само за регистриране на пикове при наблюдението на броя на слънчевите петна в периода 1700 - 2008г. (серия от 480 точки) и показва, че методът регистрира всички наблюдавани пикови стойности. Сравнен с четири други метода [7] за регистриране на пикове във времеви редици от данни се оказва, че ентропийно базирания метод работи най-добре.

МЕТОД И АЛГОРИТЪМ

Нека разгледаме N последователни във времето измервания $T = \{x_1, x_2, \dots, x_N\}$ на LAI. За да определим дали точката x_i е пик, разглеждаме локален прозорец $T_{ik} = \{x_{i-k}, \dots, x_{i-1}, x_i, x_{i+1}, \dots, x_{i+k}\}$, симетричен около точката x_i с ширина $(2k+1)$, $0 < k < i < (N-k)$. Нека дефинираме множеството от точки $T_{ik}^{(-)} = T_{ik} \setminus \{x_i\} = \{x_{i-k}, \dots, x_{i-1}, x_{i+1}, \dots, x_{i+k}\}$, т.е. множеството T_{ik} без елемента x_i , който ние искаме да оценим като пикова стойност. Като мярка за влиянието/значимостта на x_i в разглеждания прозорец, въвеждаме функцията [7]

$$S_i = H(T_{ik}) - H_{(-)}(T_{ik}^{(-)}) \quad (1)$$

където: $H(T_{ik})$ е ентропията в прозореца T_{ik} , и $H_{(-)}(T_{ik}^{(-)})$ – ентропията в прозореца $T_{ik}^{(-)}$ (T_{ik} без x_i) т.е.:

$$H(T_{ik}) = \sum_{m=i-k}^{i+k} [-p(x_i) \log p(x_i)] \quad (1a)$$

$$H_{(-)}(T_{ik}^{(-)}) = \sum_{\substack{m=i-k \\ m \neq i}}^{i+k} [-p_{(-)}(x_i) \log p_{(-)}(x_i)] \quad (1b)$$

Ако x_i е пикова стойност, тя би довела до изменение на ентропията в прозореца и тогава $S_i \geq \theta$, където θ е подходящо дефиниран праг. В (1a) и (1b) $p(x_i)$ е плътността на вероятността в точката x_i , като $p(x)$ и $p_{(-)}(x)$ са съответно функциите на плътността на вероятността в интервалите

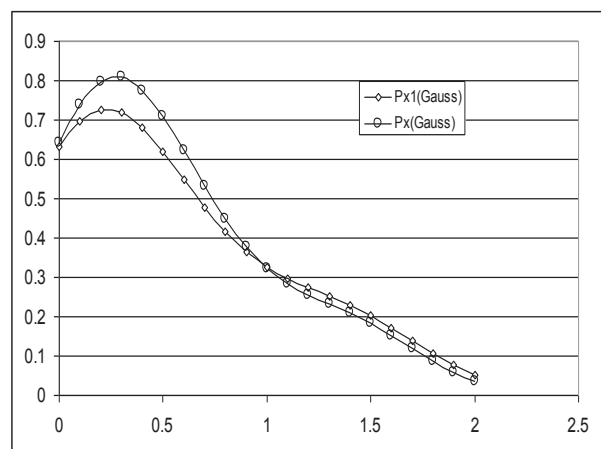
T_{ik} и $T_{ik}^{(-)}$. Плътностите на вероятността $p(x)$ и $p_{(-)}(x)$ в прозорците T_{ik} и $T_{ik}^{(-)}$ са определени посредством прозорец на Parzen-Rosenblatt [8]:

$$p_h(x) = \frac{1}{(2k+1)h} \sum_{\substack{x_j \in T \\ \text{or } T_{(-)}}} K\left(\frac{x-x_j}{h}\right) \quad (2)$$

където K е ядрото, такова че: $\int K(x)dx = 1$ and $K(x) \geq 0$, за всяко x . h е ширината на прозореца, наричана още изглаждащ параметър или ширина на честотна лента. На практика в качеството на ядрото K се взема унимодална, симетрична спрямо нулата плътност на вероятността. В настоящата работа сме използвали стандартното нормално разпределение

$$K(x) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2}x^2\right) \quad (4)$$

На фиг.1 е показан ефектът на пиковата стойност върху плътността на вероятността $p(x)$ в прозореца около нея.



Фиг.1: Влияние на точката x_i (946, 938) върху $p(x)$ функцията на плътността на вероятността (pdf) в локалния прозорец T_{ik} . P_x – pdf във T_{ik} , P_{x1} – pdf във $T_{ik}^{(-)}$. DOY = 109, 2010 (DOY – пореден ден от началото на годината)

Изборът на ширината на прозореца h е обстойно дискутиран в литературата [9]. Оптималната стойност на h се определя, като ширината на честотната лента, която минимизира интеграла на квадратичната грешка. В случая на Гаусово ядро (4), Silverman (1986) [10] дава следната оценка

$$h_{opt} \approx 0.9 \min(\sigma, \frac{IRQ}{1.34}) n^{-1/5} \quad (5)$$

$n=2k+1$ (T ширината на T прозореца), σ е стандартното отклонение и IRQ е мярка за степента на групиране на данните и е равна на разликата между 75-тия ($Q3$) и 25-тия ($Q1$) персентиля, $IRQ=Q3-Q1$.

ИЗСЛЕДВАНИЯ С РЕАЛНИ ДАННИ

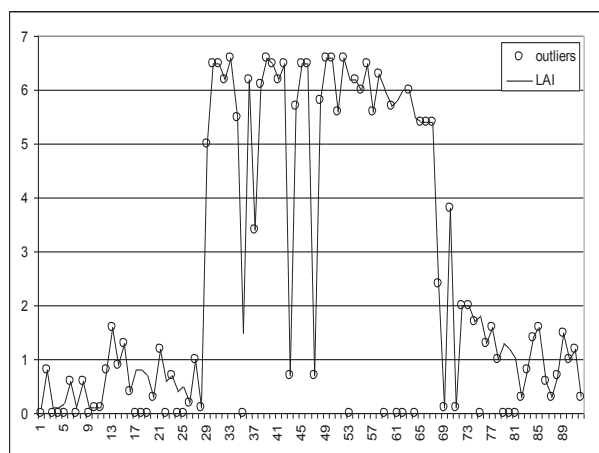
Методът, описан по-горе, е приложен за идентифициране и анализ на пиковите във времевите серии от данни LAI/FPAR (ESDT: MOD15A3), level-4. Това са комбинирани данни от измерванията със спектрорадиометъра MODIS на сателитите Terra и Aqua (NASA), композирани на всеки четири дни от годината с 1km пространствена резолюция. Продуктът LAI/FPAR съдържа измервания на индекса на листната площ (LAI) и частта на погълната фотосинтетично активна слънчева радиация (FPAR), слоеве със флагове за качеството на измерванията и оценка на стандартното отклонение при измерванията за всяка променлива. Данните са архивирани в специален формат HDF-EOS, произведен на Hierarchical Data Format (HDF),

поддържан от National Center for Supercomputer Applications (NCSA) in University of Illinois, Champaign/Urbana [11].

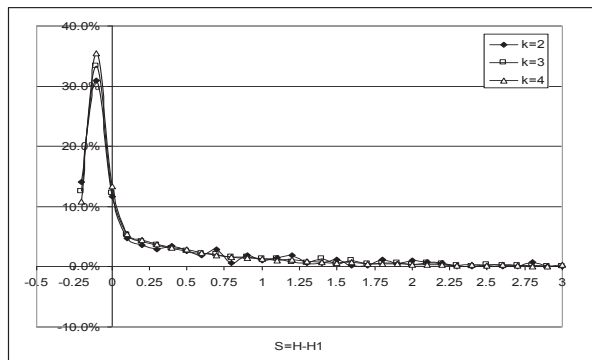
Както посочихме, предмет на настоящата работа е индексът на листната площ LAI. Анализирани са 47656 измервания на LAI, които отразяват състоянието на растителната покривка върху 518 пиксели на всеки четири дни от годината (92 стойности на LAI на година за пиксел). Пикселите включват наблюдения върху над 50000 ха широколистна гора на територията на Родопите през 2010 г.

РЕЗУЛТАТИ И АНАЛИЗ

Първоначалният анализ на стойностите на функция S , дефинирана посредством ентропията на сигнала в околностите на пиковата стойност, показва че методът успява да регистрира почти всички пикови стойности на LAI (фиг.2).



Фиг.2: Регистрирани пикови стойности на LAI ($S > 0$) за отделен пиксел, на всеки 4 дена от годината (92 измервания)



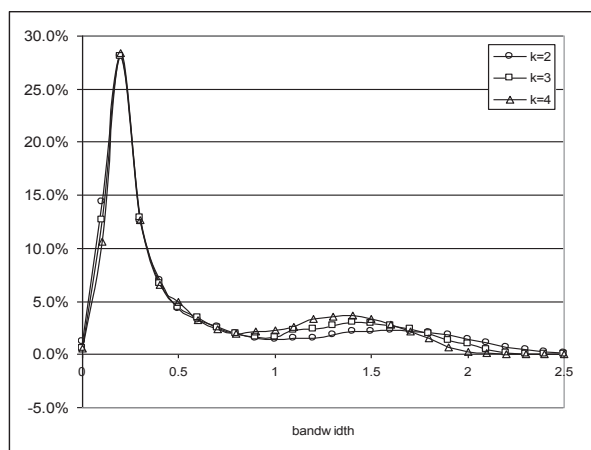
Фиг.3: Разпределение на стойностите на функцията $S = H - H_c$ през 2010 г. (47656 точки = 518 пиксела x 92 стойности на LAI за година)

На фиг.3 е показано честотното разпределение на функцията $S = H - H_c$ при различна ширина на прозореца T_{ik} (T window) около дадена точка x_i от времевата серия. Това разпределение ни показва как зависи разликата в ентропите S от стойността на праговата стойност θ . При $\theta = 0$, 42% от всички точки във времевата серия са идентифицирани като пикове. Основната част от пиковите стойности на LAI се регистрират при стойности на $\theta \in [0; 3]$. С увеличаването на прага, както може да се очаква броят на регистрираните пикове намалява, като например при $\theta = 0.1$ процентът на пиковите в LAI времевите серии е 37% от всички точки

С увеличаване на ширината на T прозореца ($n=2k+1$), нараства и броя на точките, регистрирани като пикови - 56.6% при $k=2$, 58.1% при $k=3$ и 59.7% при $k=4$.

Изследвани са и разпределенията на ширината на честотната лента h за всеки пиксел, в зависимост от ширината на T -прозореца. Разгледани са случаите на прозорци със ширина 5, 7 и 9 точки около пиковата стойност x_i (Фиг. 4). Детайлният анализ показва, че по-малки стойности на h ($h < 0.2$) преобладават при по-широките T -прозорци. Трябва да се отбележи, че това важи за преобладаващата част точки от времевите серии. Тази тенденция се обръща при h , изменящо се от 1 до 1.6 и отново се връща в интервала от 1.8 до 2.4.

Зависимостта на S от k ($n=2k+1$) е добре изразена за точките от времевите серии, за които $S \leq 0$, като там S намалява с увеличаването на k .



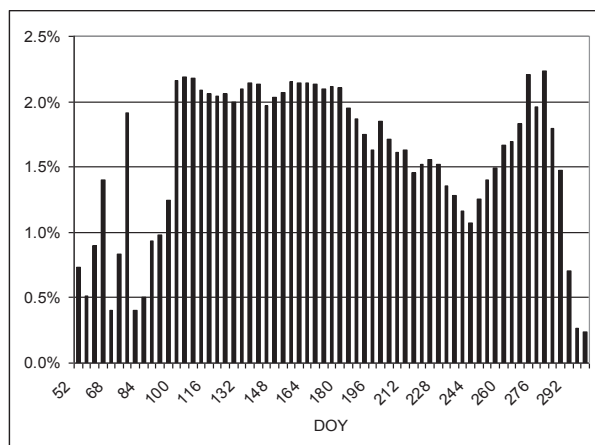
Фиг.4: Разпределение на ширината на честотната лента h в зависимост от ширината на T -прозореца ($n=2k+1$)

За изясняване на природата на пиковите е необходимо да се изследва връзката между пиковите и величини, свързани с наблюденията върху развитието на растителната покривка. От значение е да се установи разпределението на пиковите стойности през активния период на развитие на РП и връзката на пиковите със флагите за контрол, свързани с всяко измерване на LAI.

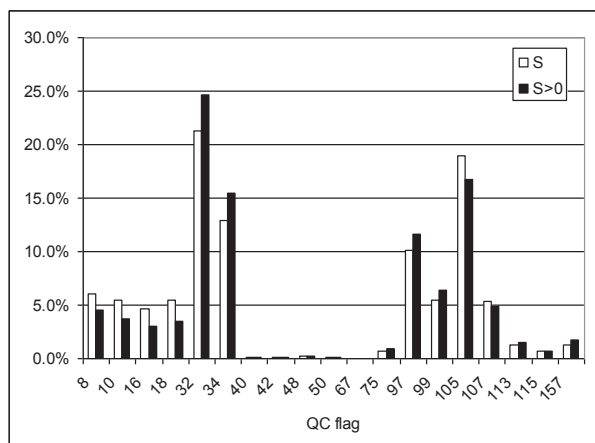
Разпределението на пиковите стойности през периода на активна вегетация е показано на фиг.5. В периода на активно натрупване на биомаса (растеж - средата на м.април до края на м.юни.), процентът на пиковите стойности е най-голям, около 2% на ден от общия брой пикове за периода. След като РП навлезе в състояние на зрялост (maturity - м. юли - м.август), броят на пиковите на ден намалява наполовина. По всяка вероятност това се дължи на стабилизирането на времето и увеличаването на дните без облаци.

За да проверим дали предложеният метод работи добре, сме анализирали връзката между регистрираните като пикови стойности на LAI и стойностите на флагите за качество (QC flags), съпровождащи всеки файл с данни от дистанционните наблюдения. Флагите за контрол на качеството на данните са разположени в 2 слоя: FparLai_QC и FparExtra_QC [12]. За настоящия анализ сме използвали данните от FparLai_QC слой, съдържащ информация за състоянието на детекторите, облачната покривка над наблюдавания регион (significant clouds NOT present-clear, significant clouds WERE present, mixed cloud present in pixel, cloud state not defined, assumed clear) и резултатите

от метода Main(RT), използван за идентифициране на стойностите на LAI. На фигура 6 е представено разпределението на пиковите в зависимост от флагите за качество (QC flags), съпровождащи измерванията на индекса на листната площ. При $QC \in \{8, 10, 16, 18\}$ разработеният метод дава леко занижена оценка на броя на наблюденията, регистрирани като пикове. Тези пиксели не са регистрирани като пикове, въпреки наличието на Significant clouds или Mixed cloud in pixel. Това се дължи на факта, че тези пиксели са с високо качество (Good quality) поради доброто сработване на Main(RT) метода, позволяващо правилното идентифициране на LAI. Основната причина за регистриране на пикселите с $QC \in \{32, 34\}$ като пикове, въпреки отсъствието на облачност е попадането на целевата функция в област на насищане.



Фиг.5: Разпределение на регистрираните пикове на LAI през вегетационния период (март – октомври, 2010), DOY – пореден ден от началото на годината



Фиг.6: Разпределение на пиковите в LAI в зависимост от флагите за качество (QC – флаг за качество [12])

Предложеният метод регистрира като пикове всички пиксели, в които LAI е определено посредством back-up алгоритъм или са запълнени въз основа на look-up таблици. В този случай, причините за по-ниското качество на данните са комплексни и включват влиянието на няколко фактора: наличието на облачност, попадане на основния алгоритъм в област на насищане или несработване на основния алгоритъм за извличане на LAI от измерените спектрални коефициенти на отражение. Методът регистрира като пикове всички случаи на излезли от строя детектори, при което повече от 50% от наблюденията са невалидни (флагова стойност $QC = \{157\}$). В такива случаи системата не извлича стойност на LAI за дадения пиксел.

ИЗВОДИ

В представената работа сме приложили успешно метода на изменение на ентропията в околностите на пикови стойности - при анализ и регистриране на пиковите в сериите от наблюдения на LAI за оценка на състоянието на РП в широколистните гори на Родопите.

Изследвана е ефективността на метода върху серия от повече от 45000 наблюдения в зависимост от ширината на Т-прозореца около пика, от праговата стойност θ , от ширината на честотната лента h . Установена е силна зависимост между флагите за качество, съпровождащи продукта LAI/FPAR от спектрорадиометъра MODIS (NASA) и изменението на ентропията в околностите на пиковите стойности на LAI в периода на активна вегетация.

Получените резултати са основа за по-пълно проучване свойствата на метода и възможностите му за приложение при изглаждане на времеви серии от измервания за установяване на състоянието на РП в глобален мащаб.

ЛИТЕРАТУРА

1. Ross, J., The Radiation Regime and Architecture of Plant Stands, Dr. W Junk, The Hague, 1980.
2. Jonsson P. and L. Eklundh, "TIMESAT - A program for analyzing time-series of satellite sensor data," Computers & Geosciences, vol. 30, no. 8, pp. 833–845, Oct. 2004
3. Cleveland, R.B., Cleveland, W.S., McRae, J.E., and Terpenning, I., 1990. STL: A Seasonal-Trend Decomposition Procedure Based on Loess. J. of Official Statistics, 6, 3-73.
4. Xiao, ZQ, Liang, SL, Wang, JD, Jiang, B, Li, XJ (2011), Real-time retrieval of Leaf Area Index from MODIS time series data, REMOTE SENSING OF ENVIRONMENT, 115(1), 97-106.
5. Hird J. N., G. J. McDermid. Noise reduction of NDVI time series: An empirical comparison of selected techniques, Remote Sensing of Environment 113 (2009) 248–258
6. Kandasamy S., F. Baret, A. Verger, P. Neveux, and M.Weiss.A comparison of methods for smoothing and gap filling time series of remote sensing observations – application to MODIS LAI products, Biogeosciences, 10, 4055–4071, 2013.
7. Palshikar G.K., Simple Algorithms for Peak Detection in Time-Series, in Proc. 1st Int. Conf. Advanced Data Analysis, Business Analytics and Intelligence (ICADABAI2009), Ahmedabad, 6-7 June 2009
8. Parzen E., (1962). On Estimation of a Probability Density Function and Mode, Annals Mathematical Statistics, Volume 33, Number 3 (1962), 1065-1076.
9. Wand, M. P. and Jones, M. C. (1995), Kernel Smoothing. London: Chapman and Hall
10. Silverman, B. W. (1986). Density Estimation for Statistics and Data Analysis, London: Chapman and Hall.
11. The HDF Group, 2011, HDF5 User's Guide, HDF5 Release 1.8.7, May 2011
12. MODIS LAI/FPAR Product User's Guide (May 21, 2012), https://lpdaac.usgs.gov/products/modis_products_table/mcd15a3