OPTIMIZING LUT-BASED RADIATIVE TRANSFER MODEL INVERSION FOR RETRIEVAL OF BIOPHYSICAL PARAMETERS USING HYPERSPECTRAL DATA

J. Verrelst, G.P. Rivera, G. Leonenko, L. Alonso and J. Moreno

Image Processing Laboratory (IPL). Universitat de València. E-mail: jochem.verrelst@uv.es

ABSTRACT

Inversion of radiative transfer models using a lookup-table (LUT) approach against hyperspectral data streams leads to retrievals of biophysical parameters such as chlorophyll content (Chl), but necessary optimization strategies are not consolidated yet. Here, various regularization options have been evaluated to the benefit of improved Chl retrieval from hyperspectral CHRIS data, being: i) the role of added noise, ii) the role of multiple best solutions, and iii) the role of applied cost functions in LUT-based inversion. By using data from the ESA-led field campaign SPARC (Barrax, Spain), it was found that introducing noise and opting for multiple best solutions in the inversion considerably improved retrievals. However, the widely used RMSE was not the best performing cost function. Three families of alternative cost functions were applied here: information measures, minimum contrast and M-estimates. We found that so-called 'Power divergence measure', 'Trigonometric' and spectral measure with 'Contrast function K(x)=-log(x)+x' outperformed RMSE. The whole inversion approach, including more than 60 different cost functions, has been implemented in the ARTMO (Automated Radiative Transfer Models Operator) GUI toolbox and can easily be applied to other kinds of multispectral or hyperspectral images.

***Index Terms*—** LUT-based inversion, Chlorophyll content retrieval, cost functions, radiative transfer models, CHRIS

1. INTRODUCTION

Leaf chlorophyll content (Chl) is among the most important biophysical parameters retrievable from satellite reflectance data. The parameter gives insight in the phenological stage and health status (e.g., development, productivity, stress) of crops and forests.

When it comes to the development of retrieval methods, it is mandatory to invest in methods that are both accurate and robust and at the same time can be applied in an operational context. Contrary to empirical approaches, canopy radiative transfer models (RTMs) explicitly interpret driving processes between solar radiation and the elements constituting the canopy using physical laws. Because of being physically-based, inversion of canopy RTMs against actual EO data is generally considered as one of the most accurate approaches to map biophysical parameters. However, this approach is not straightforward. According to Hadamarad postulates, mathematical models of physical phenomena are mathematically invertible if the solution of the inverse problem to be solved exists, is unique and depends continuously on variables. Unfortunately this assumption is not met. In fact, the inversion of canopy RTMs is by nature an ill-posed problem mainly for two reasons: on the one hand, several combinations of canopy biophysical and leaf biochemical parameters have a mutually compensating effect on canopy reflectance thus leading to very similar solutions. On the other hand, model uncertainties and simplifications (e.g. 1D nature of some models) may induce large inaccuracies in the modelled canopy reflectance.

Over the past two decades, different successful strategies have been proposed to circumvent the drawback of ill-posedness, with Lookup-table (LUT)-based inversion strategies as most popular ones. The main advantage of LUT-based inversion approaches is that it can be fast because the most computationally expensive part of the inversion procedure is completed before the inversion itself. LUT-based inversion in its essential form, i.e. direct comparison of LUT spectra against an observed spectra through a cost function, constitutes the majority of applied inversion approaches. Various regularization strategies have been proposed to increase the robustness of the estimates: 1) the use of prior knowledge about model parameters (1), 2) the use of multiple best solutions (instead of the single best solution) (1), 3) adding noise to account for uncertainties attached to measurements and models (2), and, 4) the combination of single variables into synthetic variables such as the canopy level content of absorbing materials (3), e.g. canopy Chl, which is the product of leaf Chl and LAI.

Nevertheless, in view of applying these regularization strategies into a more operational context, above-mentioned studies are constrained in various ways. First, while the majority of reviewed studies focused on optimizing a single LUT-based inversion problem, the mutual impact of proposed optimizing strategies has not been systematically assessed. Second, in each of these studies the well-known root mean square error (RMSE) was used as cost function between simulated and measured spectra. However, in case of outliers and nonlinearity, the residuals are distorted and therefore the key assumption for using RMSE (Gaussian or zero mean white noise distribution of residuals) is violated (4). The latter authors suggested that alternative cost functions may provide a more robust way to estimate biophysical parameters since they allow retrievals for cases where errors are not normally distributed and allow to deal with nonlinear high-parametric problems. The availability of a large number of cost functions gives a high degree of flexibility, since it allows model optimization for a wide range of error distributions. Hence, alternative cost functions deserve to be evaluated in view of

J. Verrelst is supported by the FP7-PEOPLE-IEF-2009 grant (Grant Agreement 252237).

the above-described optimizing strategies. Third, the majority of these studies focus on a specific vegetation type such as crop types, identified within the image. This assumes that up-to-date knowledge of land cover types is available, which is usually not the case in an operational context. Moreover, eventually LUT-based inversion should be applicable not only for agroecosystems but over all natural and semi-natural vegetated surfaces. And fourth, from a practical perspective, hardly any of aforementioned studies provide links to software packages where proposed optimization strategies have been implemented. In this work we aim to systematically evaluate different LUT-based inversion strategies in view of hyperspectral data for the benefit of improved pixel-wise estimation ofChl. The mutual impact of the following strategies were investigated: 1) the role of added noise, 2) the role of multiple best solutions, and finally: 3) the role of applied cost functions in these strategies. Data used came from the ESA-led field campaign SPARC, which took place on the agricultural test site Barrax, Spain.

2. ARTMO GUI TOOLBOX

In an attempt to automate LUT-based inversion, the whole processing chain has been implemented into ARTMO (Automated Radiative Transfer Models Operator) (5). ARTMO is a GUI toolbox written in Matlab in which the user can choose from multiple leaf and canopy RT models to generate class-based LUTs. This innovative toolbox provides essential tools for running and inverting a suite of plant reflectance models. In short, the toolbox enables the user: i) to choose between various plant leaf (e.g., PROSPECT-4, PROSPECT-5) and canopy reflectance models (e.g., 4SAIL, SLC, FLIGHT), ii) to choose between spectral band settings of various air- and space-borne sensors or defining new sensor band settings, iii) to simulate a massive amount of spectra and storing them in a relational database, iv) to visualize spectra of multiple models in the same plotting window, and finally, v) to run LUT-based model inversion against EO imagery given a selected cost function, optimization options and accuracy estimates. Moreover, ARTMO is able to run inversions per land cover class, which permits realistic retrievals of biophysical parameters over patchy landscapes. For instance, agricultural fields can be interpreted by a 1D model while forests can be interpreted by a 3D model.

3. METHODOLOGY

3.1. Cost functions

Numerical solution of the inverse problem adjusts the model parameters such that model predicted values closely match the measured values. The match between model output and data is usually based on minimizing the sum of least squares, as in RMSE. Another way to obtain better estimates is using alternative cost functions, e.g. as those introduced in (4). The latter authors investigated several families of cost functions on simulated reflectance data for conifer and broadleaf cover. In general, statistical distances can be categorized into three families: *information measures*, *minimum contrast* and *M-estimates*. Although they all represent 'distance' or 'metric' between two functions the main difference of these families is the way how reflectance functions are interpreted and in what space. These metric families came from different areas of mathematics and statistics and play an important role in image processing, engineering, medicine and code theory. They allow to take into the account nonlinearity of the problem, robustness and skewness of the noise to provide better retrievals of biophysical parameters. Moreover, contrary to using one cost function, the availability of a large number of statistical distances or divergence measures gives a high degree of flexibility, since it allows model optimization for different assumptions on the nature and properties of errors.

Based on the research in (4) we initially compared all the available cost functions (62) and selected three cost functions from the different families which may provide promising results as alternative to RSME. Let $D[P,Q]$ represent a distance between two functions, where $P = (p(\lambda_1), \cdots, p(\lambda_n))$ is satellite and $Q = (q(\lambda_1), \cdots, q(\lambda_n))$ is LUT correspondent reflectances and $\lambda_1, \cdots, \lambda_n$ represent n bands. Thus, for RMSE:

$$D[P,Q] = \sqrt{\sum_{\lambda_i=1}^{\lambda_n} \frac{(p(\lambda_i) - q(\lambda_i))}{n}} \qquad (1)$$

First alternative cost function belongs to the family of *information measures*. This class represents different distances between two probability distributions and were widely explored throughout mathematical applications. In this case we consider reflectance as probability distribution function and normalization is required (sum of probabilities is 1) prior to numerical application. Within this family, the 'Power divergence measure' has the following form:

$$D[P,Q] = \sum_{\lambda_i=1}^{\lambda_n} p(\lambda_i) \frac{\{[p(\lambda_i)/q(\lambda_i)]^\alpha - 1\}}{\alpha(\alpha+1)}, \alpha \in (-\infty, +\infty) \qquad (2)$$

Note that in some cases for parameter $\alpha = -2, -1, -1/2, 0, 1$ we can get the following already known measures: the Neyman chi-squared measure divided by 2, the Kullback-Leibler divergence, the twice-squared Hellinger distance, the likelihood disparity, and the Pearson's chi-squared divided by 2.

Second alternative cost function belongs to the family of *M-estimates*. In this case we interpret reflectance as nonlinear regression function. One of the well known distances from this class is RMSE which, for Gaussian error distributions, is consistent, asymptotically normal and asymptotically efficient. However, when the error distribution is non-Gaussian or non-symmetric, the RMSE can result in large losses of efficiency. Robust methods replace the sum of squares by more suitable loss functions. Thus, for $\alpha, \beta > 0$, we can determine so-called 'Trigonometric' distance:

$$D[P,Q] = \sum_{\lambda_i=1}^{\lambda_n} \alpha x(\lambda_i) arctan(\beta * x(\lambda_i)) - \alpha \frac{log(s^2(x(\lambda_i))^2 + 1)}{2\beta}, \qquad (3)$$

where $x(\lambda_i) = p(\lambda_i) - q(\lambda_i)$. It is known that errors in this case are distributed by a logistic distribution.

Third alternative cost function belongs to the family of *minimum contrast estimates*, where we consider reflectance as a spectral density function of some stochastic process. The basic idea behind it is to minimize the distance (contrast) between a parametric model and a non-parametric spectral density. Since one can interpret satellite observations as measurements in the spectral domain these distances seem to be a natural choice for analysing satellite data. We consider the following spectral distance with the so-called 'Contrast function $K(x) = -log(x) + x$', then distance has the form:

$$D[P,Q] = \sum_{\lambda_i}^{\lambda_n} \{-log(q(\lambda_i))/p(\lambda_i)) + q(\lambda_i))//p(\lambda_i)\}. \quad (4)$$

3.2. SPARC database

Ideally, LUT-based inversion strategies should be validated by a dataset that represents the same variety of actual crops and conditions as remotely observed by the optical sensor. A diverse field dataset, covering various crop types, growing phases, canopy geometries and soil conditions was collected during SPARC (SPectra bARrax Campaign). The SPARC-2003 and SPARC-2004 campaigns took place in Barrax, La Mancha, Spain (coordinates 30°3'N, 28°6'W, 700 m altitude). The test area has an extent of 5 km × 10 km, and is characterized by a flat morphology and large, uniform land-use units. The region consists of approximately 65% dry land and 35% irrigated land. The annual rainfall average is about 400 mm. In the 2003 campaign (12-14 July) biophysical parameters were measured within a total of 113 Elementary Sampling Units (ESU) among different crops. ESU refers to a plot size of about 20^2 m^2. The same field data were collected in the 2004 campaign (15-16 July) within a total of 18 ESUs among different crops. Leaf Chl was derived by measuring within each ESU about 50 samples with a calibrated CCM-200 Chlorophyll Content Meter. For both years, we have a total of 9 crops (garlic, alfalfa, onion, sunflower, corn, potato, sugar beet, vineyard and wheat), with field-measured values of LAI that vary between 0.4 and 6.3 and Chl between 2 and 55 μg/cm^{-2}. Further details on the measurements can be found in (6).

3.3. LUT generation

From the available models in ARTMO we chose to use the coupling between PROSPECT-4 and 4SAIL because of being fast, invertible and well representing homogeneous plant covers on flat surfaces areas such as those present at Barrax. Both models, hereafter referred as PROSAIL, have been used extensively over the past few years for a variety of applications (for a review on these models see (7)). PROSPECT-4 calculates leaf reflectance and transmittance over the solar spectrum from 400 to 2500 nm at a 1 nm spectral sampling interval as a function of its biochemistry and anatomical structure. It consists of 4 parameters, being leaf structure, chlorophyll content (Chl), equivalent water thickness and dry matter content. 4SAIL calculates top-of-canopy reflectance. 4SAIL inputs consist of: LAI, leaf angle distribution, ratio diffuse/direct irradiation, a hot spot parameter and sun-target-sensor geometry. Spectral input consists of leaf reflectance and transmittance spectra, here coming from PROSPECT-4, and a a moist and dry soil reflectance spectrum. To obtain these soil spectra, the average of bare soil signature was calculated from bare moist and dry soil pixels identified in the imagery. In 4SAIL a scaling factor, α_{soil}, has been introduced that takes variation in soil brightness into account as a function of these two soil types.

The bounds and distributions of the PROSAIL variables are defined according to (8). They were chosen in order to describe the characteristics of all crop types used in the study. Gaussian input distributions were generated for Chl content in order to put more emphasis on the variable values being present in the actual growth stages of the crops. Sun and viewing conditions correspond to the situation of the satellite overpass. A useful feature of ARTMO is that all simulations are automatically employed according to predefined band settings (here CHRIS mode 1). From the inserted leaf and canopy input ranges all possible combinations were calculated by ARTMO. From this, a LUT size of 100000 TOC reflectance realizations was randomly chosen. All input parameters, metadata and associated output simulations were automatically stored in a relational database running underneath ARTMO.

Two regularization options are commonly applied in LUT-based inversion strategies. First, often a Gaussian (white) noise is added to the simulated canopy reflectance. Different numbers are encountered in literature, typically spanning from 2.5 to 20% (1), meaning that this strategy is not consolidated yet. To clarify its role in LUT-based inversion, a systematic assessment is pursued here, ranging from 0 (no noise) until 30% noise. Second, several studies demonstrated that the single best parameter combination corresponding to the smallest RMSE does not necessarily lead to best accuracies (1). A widely applied strategy is therefore taking the mean of multiple best solutions. Also here different numbers are encountered in literature, spanning from the single best solution to the mean of the 20% best solutions. However, its role in view of different noise levels and cost functions remains to be evaluated. Therefore, a range from 0 (single best solution) to the mean of 30% best solutions has been included in the analysis. Given all these factors, their effects on the robustness of LUT-based inversion has been assessed. The retrieved predictions were compared against the measured validation dataset using the normalized or relative RMSE (RRMSE).

4. NUMERICAL RESULTS

The performance of the four cost functions in retrieving Chl validated against the field dataset using the relative RMSE (RRMSE); further referred as relative error to avoid confusion with RMSE as cost function. Figure 1 shows relative error matrices displaying the impact of noise levels against multiple best solutions. Several observations can be made from these error matrices. First, the widely used RMSE was not evaluated as best performing cost function (best result: 26.30%). All the alternative functions outperformed RMSE with their best results. Particularly 'Power divergence' proved to be a more robust cost function, with best relative error obtained at 18.93%, followed by 'Trigonometric' (22.33%) and then 'Contrast function K(x)=-log(x)+x' (26.09%). Second, for none of the cost functions was the single best solution and without noise evaluated as best configuration. In fact, adding noise and applying multiple best solutions in the inversion considerably improved accuracies, e.g. for RMSE relative errors lowered from 43.64% (no regularization options) to 26.30% (30% best solutions, 11% noise added). Also for the others accuracies improved spectacularly (e.g. Trigonometric from 52.67% to 22.33%), which underlines the importance of regularization strategies. It should however be noted that 'Power divergence' needs one parameter and 'Trigonometric' even two parameters to be tuned. Given these four examples it can be concluded that RMSE behaves as suboptimal cost function. It should also be mentioned that in ARTMO the best evaluated cost function and regularization options can immediately be applied over the whole image so that maps of the biophysical parameters can be generated in an automated way.

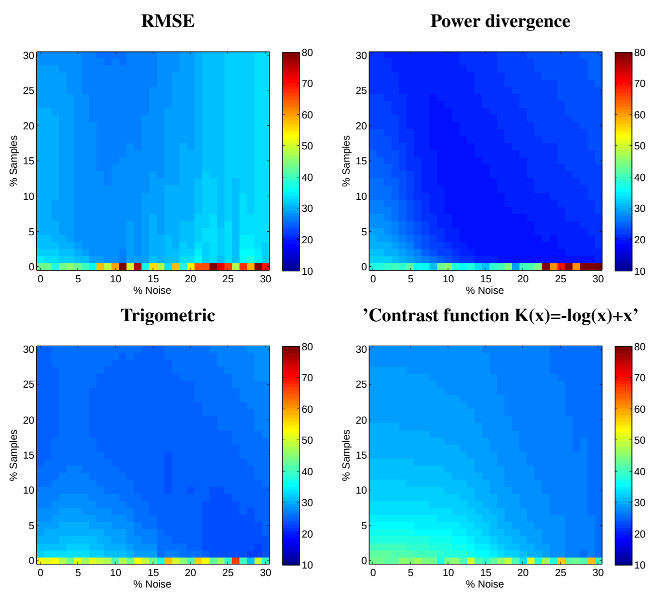

Fig. 1. Relative RMSE (RRMSE) matrices using 'RMSE', 'Power divergence', 'Trigonometric' and 'Contrast function K(x)=-log(x)+x' as cost functions displaying the impact of % noise (X-axis) vs. multiple solutions (Y-axis) in LUT-based RTM inversion for Chl retrieval against CHRIS observations. The more bluish, the better the estimate.

5. CONCLUSIONS

While various LUT-based inversion methods have been proposed in literature for retrieval of biophysical parameters they all rely on RMSE as cost function. However RMSE can result in large losses of efficiency when the error distribution is non-Gaussian or non-symmetric. For the benefit of realizing improved retrievals, we have compared three alternative cost functions, being 'Power divergence measure', Trigonometric' and 'Contrast function K(x)=-log(x)+x') using a hyperspectral CHRIS dataset (62 bands). These cost functions outperformed the widely used RMSE in Chl retrieval. Introducing noise and applying multiple best solutions in the inversion further improved the inversion performance. The whole inversion processing chain, including more than 60 different cost functions and regularization options, has been implemented in the ARTMO toolbox. Specifically, an 'inversion evaluation' module has been implemented that enables comparing all available cost functions and regularization options against a test dataset prior to applying the inversion over the whole image. ARTMO is available on request to those who want to contribute in expanding the toolbox.

References

[1] B. Combal, F. Baret, M. Weiss, A. Trubuil, D. Mac, A. Pragnee, R. Myneni, Y. Knyazikhin, and L. Wang, "Retrieval of canopy biophysical variables from bidirectional reflectance using prior information to solve the ill-posed inverse problem," *Remote Sensing of Environment*, vol. 84, no. 1, pp. 1–15, 2003.

[2] W. A. Dorigo, R. Zurita-Milla, A. J. W. de Wit, J. Brazile, R. Singh, and M. E. Schaepman, "A review on reflective remote sensing and data assimilation techniques for enhanced agroecosystem modeling," *International Journal of Applied Earth Observation and Geoinformation*, vol. 9, no. 2, pp. 165–193, 2007.

[3] M. Weiss, F. Baret, R.B. Myneni, A. Pragnee, and Y. Knyazikhin, "Investigation of a model inversion technique to estimate canopy biophysical variables from spectral and directional reflectance data," *Agronomie*, vol. 20, no. 1, pp. 3–22, 2000, cited By (since 1996) 111.

[4] Leonenko G., North P.R.J., and Los S.O., "Statistical distances and their applications to biophysical parameter estimation. Information measures, m-estimates, minimum contrast method," *IEEE Transactions on Geoscience and Remote Sensing*, vol. In review, 2012.

[5] J. Verrelst, J.P. Rivera, L. Alonso, and J. Moreno, "ARTMO: an Automated Radiative Transfer Models Operator toolbox for automated retrieval of biophysical parameters through model inversion," *Proceedings of EARSeL 7th SIG-Imaging Spectroscopy Workshop, Edinburgh, UK*, 2011.

[6] J. Delegido, G. Fernandez, S. Gandía, and J. Moreno, "Retrieval of chlorophyll content and LAI of crops using hyperspectral techniques: Application to PROBA/CHRIS data," *International Journal of Remote Sensing*, vol. 29, no. 24, pp. 7107–7127, 2008.

[7] S. Jacquemoud, W. Verhoef, F. Baret, C. Bacour, P.J. Zarco-Tejada, G.P. Asner, C. Francois, and S.L. Ustin, "PROSPECT + SAIL models: A review of use for vegetation characterization," *Remote Sensing of Environment*, vol. 113, no. SUPPL. 1, pp. S56–S66, 2009.

[8] K. Richter, C. Atzberger, F. Vuolo, P. Weihs, and G. D'Urso, "Experimental assessment of the sentinel-2 band setting for RTM-based LAI retrieval of sugar beet and maize," *Canadian Journal of Remote Sensing*, vol. 35, no. 3, pp. 230–247, 2009.