Selection of PolSAR Observables for Crop Biophysical Variable Estimation With Global Sensitivity Analysis

How to cite:
Selection of PolSAR Observables for Crop Biophysical Variable Estimation with Global Sensitivity Analysis

Esra Erten, Gülşen Taşkin, Juan M. Lopez-Sanchez

Abstract—The role of global sensitivity analysis (GSA) is to quantify and rank the most influential features for biophysical variable estimation. In this paper, an approximation model, called High Dimensional Model Representation (HDMR), is utilized to develop a regression method in conjunction with a global sensitivity analysis in the context of determining key input drivers in the estimation of crop biophysical variables from polarimetric Synthetic Aperture Radar (PolSAR) data. A multitemporal Radarsat-2 dataset is used for retrieval of three biophysical variables of barley: leaf area index, normalized difference vegetation index and BBCH-stage. The HDMR technique is first adopted to estimate a regression model with all available polarimetric features for each biophysical parameter, and sensitivity indices (SI) of each feature are then derived to explain the original space with a smaller number of features in which a final regression model is established. To evaluate the applicability of this methodology, root mean square and coefficient of determination were performed under different amounts of samples. Results highlight that HDMR can be used effectively in biophysical variable estimation for not only reducing computational cost, but also for providing a robust regression.

Index Terms—Agriculture, polarimetry, synthetic aperture radar, global sensitivity analysis (GSA), Radarsat-2

I. INTRODUCTION

Tracking crop biophysical variables using time series of polarimetric Synthetic Aperture Radar (PolSAR) data is an active research topic in precision agriculture due to the sensitivity of PolSAR data to the physical and geometrical structure of vegetation [1], [2]. From the application point of view, its all-weather, day and night imaging capability makes PolSAR well-suited for tracking crop status and the evolution of its biophysical variables.

Retrieval of crop variables using PolSAR, or any other remote sensing technique, can roughly be categorized under two approaches: physical and statistical. Physical models express the backscattering from crops as a complex function of morphological parameters of crop with assumptions on ground scattering and dielectric constant [2]. Even though there are some successful examples, when it comes to operational monitoring, the requirement of a physical scattering model for each crop type limits the performance of physical model-based crop monitoring approaches. Instead, statistical approaches (specifically regression based on machine learning) are getting very popular for crop monitoring. The ease of their implementation at an affordable computational cost makes machine learning regression algorithms appealing for crop monitoring [3]–[5]. Learned or data-driven statistical models do not deal with equations of physical laws (e.g. backscattering model for a crop). Instead, they are more flexible since they define the biophysical variable estimation as a regression problem, hence rely heavily on the availability of training samples.

An essential part of the machine learning based biophysical variable estimation is the uncertainty quantification (UQ) of the regression coming from the inputs (e.g. measured PolSAR data). UQ provides an assessment of the impact of the inputs distribution on outputs distribution. UQ studies are classified into two groups: local sensitivity analysis (LSA) and global sensitivity analysis (GSA). The LSA consists of partial derivatives of model outputs with respect to the input variables, meaning that it is only capable of measuring the local effect of the input variables on the model output. Therefore, LSA does not consider the effect of the interactions between the input parameters on the model output. On the contrary, GSA allows us to measure the effect of the uncertainty of each input variable on the variance or distribution of the model outputs. High Dimensional Model Representation (HDMR) is one of the most popular methods for computing the global sensitivity indices. However, it requires analytical form of the input function in advance to calculate each component in the HDMR expansion. When the input function is not known or very complex, Random Sampling HDMR (RS-HDMR) was proposed, which is based on the decomposition of the HDMR terms with splines or orthogonal polynomials [6]. However, a significant drawback of the classical HDMR is the need of the input variables to cover uniformly the whole input space, which is not the case in many occasions due to the variability and scarcity of the input data, obtained by space-borne radar measurements. To handle such measurements, in this work, the RS-HDMR was solved in the least squares sense with a regularizer to prevent unstable behaviors of the coefficients of the orthogonal polynomials.

Sensitivity analysis has recently attracted much interest in remote sensing in the context of estimation the relevant scene features. For example, in [7], a Gaussian Process Regression (GPR) based local sensitivity analysis was introduced by making use of mean and variance predictive estimates of the GPR for global ocean chlorophyll prediction to reveal the most important spectral bands. In the context of the GSA, a comprehensive evaluation that combines physically based emission models and various global SA algorithms to evaluate the sensitivities of microwave emissivity and brightness temperature to soil parameters was presented in [8]. A GSA of a radiative transfer theory based scattering model to see how different morphological parameters of rice impact on a backscattering under given set of assumptions was performed by [2]. Compared to these studies, our work aims to expand the applicability of the GSA to a regression method
constructed by meta-modelling, which does not need to take into account the physics behind the acquisition system. The proposed method in this study is primarily based on so-called High Dimensional Model Representation Ridge Regression (HDMR-RR) as presented in [9]. Numerous studies have been conducted with the HDMR for function approximation [10] and sensitivity analysis [11], and very promising results have been obtained in different ranges of applications. The main contribution of this paper is the utilization of the HDMR-RR for a global sensitivity analysis and the demonstration of its performance for crop biophysical variable estimation with POLSAR data as input.

II. METHODOLOGY: HDMR-RR

HDMR is a method which decomposes a multivariate function with \( n \) variables into a finite hierarchical expansion of component functions with respect to input variables [12], as follows:

\[
y = f(x) = f_0 + \sum_{i=1}^{n} f_i(x_i) + \sum_{i<j} f_{ij}(x_i, x_j) + \cdots + f_{1\cdots n}(x_1, \ldots, x_n)
\]

where \( f_0 \) refers to the constant term. The component functions \( f_i(x_i) \) are called first order terms, \( f_{ij}(x_i, x_j) \)-second order terms, and so on. Each component function is calculated by taking multi-dimensional integrations, which are very computationally demanding [13]. Random Sampling HDMR (RS-HDMR) was proposed to alleviate this by approximating the component functions as a finite combination of orthogonal polynomials. The HDMR-RR is a first order regularized least squares regression method which utilizes the first order RS-HDMR terms in the approximation, as shown as below:

\[
f(x) \approx f_0 + \sum_{i=1}^{m} \sum_{j=1}^{m} \alpha_i^j \varphi_j(x_i)
\]

where \( m \) is the order of the orthogonal Legendre Polynomials (LPs), \( \alpha_i^j \) and \( \varphi_j(x_i) \) refer to the coefficient of the LP and its value at the sample \( x_i \). The constant term, mean response of \( f \), is simply calculated by taking the average of all the outputs. Having vectorized (2), the LPs coefficients can be found by solving the following regularized least squares optimization problem:

\[
\mathcal{L} = \min_{\alpha} \sum_{i=1}^{l} \|1_{\mathbb{F}} + \alpha^T \Phi(x^i) - y \|^2 + \lambda |\alpha|^2
\]

where \( \alpha \) is a vector of LPs coefficients associated with each feature, \( y \) and \( \lambda \) correspond to the model output and the Tikhonov regularization coefficient, respectively, while \( \Phi \) is a matrix obtained by explicitly evaluating the LPs at each \( x_i \) sample in the input domain:

\[
\Phi = \begin{bmatrix}
\varphi_1(x_1^i) \cdot \varphi_m(x_1^i) & \cdots & \varphi_1(x_n^i) \cdot \varphi_m(x_n^i) \\
\varphi_1(x_1^i) \cdot \varphi_m(x_1^i) & \cdots & \varphi_1(x_n^i) \cdot \varphi_m(x_n^i)
\end{bmatrix}
\]

Where \( l \) corresponds to the number of samples. By solving Eq. 3, \( \alpha \) can be obtained as below:

\[
\alpha_{m+1} = (\Phi^T \Phi + \lambda I)^{-1} \Phi^T y
\]

A. Global Sensitivity Analysis: Sobol Indices

The functional decomposition of the variance via HDMR is often referred to ANAlysis Of Variance (ANOVA), as shown in Eq.1. It is demonstrated that such a decomposition exists for every finite-variance functional, and it is orthonormal, hence yielding unique coefficients [14]. ANOVA enables global sensitivity analysis for uncorrelated input variables [15]. A widespread quantitative global sensitivity measure is given by the variance-decomposition-based Sobol indices (SI) [14].

The straightforward calculation of the SI based on the ANOVA HDMR requires \( n^2 \) integral evaluations, which is impractical in many of the engineering problem. Another way to calculate these coefficients is the use of regression [13], [16]. Therefore, in this study, the proposed approach is considered as a global sensitivity tool.

Once the LPs coefficients are determined via the HDMR-RR, the first order SI can be calculated as the ratio of the variance of each component function, \( D_i \), to the total variance \( D \), as follow:

\[
S_i = D_i/D, \quad i = 1, \ldots, n
\]

where

\[
D_i \approx \sum_{j=1}^{m} \langle x_j^i \rangle^2, \quad D = \sum_{i=1}^{n} D_i
\]

It can be easily seen that there is a close relationship between variance decomposition and HDMR-RR coefficients, which allows for the calculation of Sobol indices directly from the Legendre coefficients, i.e. without the need for additional sampling as it is required in variance decomposition of the classical HDMR. Note that first order sensitivity indices show the effect of the single factor \( x_i \) on the output \( f(x) \) but it does not account for the high dimensional component functions, in which the sum of the SIs corresponding to the all component functions is equal to one.

III. STUDY AREA AND DATASET

A dataset formed by 20 sequential Radarsat-2 images and in-situ measurements of various crop types, acquired over Indian Head (Canada) as part of the ESA-funded AgriSAR2009 campaign is exploited here. As a representative cereal, barley is considered in all the experiments of this study. As part of the AgriSAR2009 campaign data, biophysical variables such as LAI, NDVI and phenological stages were measured within 3 barley fields. The phenological scale employed in the ground campaign is the continuous BBCH (Biologische Bundesanstalt, Bundessortenamt und CHemischeIndustrie), with a range \([0 \ldots 100]\). LAI and NDVI values were obtained in-situ with the Plant Canopy analyzer LAI-2000 and the GreenSeeker RT200 Nitrogen/Defoliant/Plant Growth, respectively (see [17], [18] for more details).

For each field, the in-situ measurements were conducted weekly at four spatially independent points, yielding a total of 144, 108 and 100 samples for BBCH stage, NDVI and LAI, respectively. Fig. 1 shows the acquisition dates of Radarsat-2 and the LAI measurements of barley along with phenological information. The BBCH value increases monotonically in time whereas the LAI, a dimensionless quantity defined as the ratio of the total area of all leaves on a plant to ground surface.
TABLE I
LIST OF RADARSAT-2 POLARIMETRIC FEATURES [1/4] AND CROP VARIABLES

<table>
<thead>
<tr>
<th>Polarimetric features</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>f1, f2, f3, f4, f5, f6, f7</td>
<td>$</td>
</tr>
<tr>
<td>f8, f9, f10</td>
<td>$a_{HH}, a_{VV}, a_{HH-VV}$</td>
</tr>
<tr>
<td>f11, f12, f13</td>
<td>$\phi_{HH,HH}, \phi_{HH,VV}, \phi_{HH,V-H}$</td>
</tr>
<tr>
<td>f14, f15, f16</td>
<td>$\phi_{VV,HH}, \phi_{VV,HH}, \phi_{HH,VH}$</td>
</tr>
<tr>
<td>f17, f18, f19, f20</td>
<td>$</td>
</tr>
<tr>
<td>f21</td>
<td>$\rho_{HH,HH}, \rho_{HH,VV}$</td>
</tr>
</tbody>
</table>

Biophysical variables Description
---
LAI | green leaf area per unit
NDVI | difference between reflected near-infrared and red light
Phenology | BBCH scale (0-100)

According to the results, the HDMR provides slightly better accuracy in terms of RMS and R2 than the other regression methods on the LAI dataset. However, for NDVI and BBCH estimations, although the GPR outperforms the others in terms of the median value of the metrics, its mean accuracy is not statistically different than the HDMR, which has comparable stability as well. For all the estimations, in terms of the median value, the HDMR-RR obtains R2 values greater than 0.8 (a one value would indicate a perfect prediction of the outputs from the inputs). However, the small sample size (see Fig. 1) and the evident large variations in LAI within the fields, make the regression more difficult to perform for LAI than for the other two variables. In a nut shell, the results show that the HDMR-RR achieves a performance competitive with the other methods and shows superiority because of the global sensitivity analysis.

In order to determine the most relevant features for each variable estimation, the sensitivity indices were computed using (6) based on the coefficient vector $\alpha$. Fig. 3 shows the results of the sensitivity indices for BBCH, NDVI, and LAI as a heat map for each trial. The highest values indicate the features most relevant for improving the performance of the regression. Note that compared to LAI and NDVI, the most important feature for BBCH has a SI around 0.3 meaning that there is no single dominant feature contributing to the BBCH variability. The main reason of this, based on the BBCH modeling, is that more plant features affect the definition of BBCH than the definition of LAI and NDVI. Phenology, i.e. BBCH stage, is determined by the morphological, chemical and physical condition of the field. Instead, LAI is a function of physical structure of the crops, hence, being able to be properly characterized by less number of polarimetric features.

Regarding the relevance of the individual features, $f_8$, the co-polar backscattering ratio, is the most important feature for all the biophysical variables, as shown in Fig. 3. The importance of $f_8$ comes from the crops physical condition in time, producing different attenuation behaviour for vertical and horizontal polarisation. Indeed, the change in size and
vertical structure of barley makes the co-polar backscattering ratio the most relevant feature among the others. According to Fig. 3(a), in addition to the co-polar backscattering ratio, the phase differences between co-polar (f14) and 1st and 2nd Pauli channels (f20) and \( \alpha \) (f7) show high values of sensitivity indices, meaning that the features related to the scattering mechanism are outstanding features among the others listed in Table I for the BBCH stage of barley. For LAI and NDVI, in addition to \( \beta \), the correlation of co-polar channels and 1st and 2nd Pauli channels are the most prominent feature. Despite \( \alpha \) (f7) is an influential feature in BBCH scale estimation, it can be ignored in the estimation of the NDVI and the LAI. Instead, entropy (degree of randomness of scattering) appears important within the features. It is worth noting that here only the first-order SI-based on the uncorrelated features assumption- are considered. The second- and higher-order SI expressing the interactions among the features are not considered. This is the reason highly correlated two features may not have exactly same SI, but they have similar SI, hence similar importance.

In order to analyse the effects of the relevant features in terms of \( R^2 \) and RMS, having ranked the features with respect to their highest SI, averaged over one hundred trials, the relevant features were given to the HDMR-RR one at a time. Fig. 4 illustrates how the number of ranked features at each run of biophysical variable estimation affects the prediction results, measuring both \( R^2 \) and RMS. It can be easily seen that the HDMR-RR captures the peak accuracy when the number of features are 5, 6 and 9 for NDVI, LAI and BBCH, respectively. After reaching the peak accuracy, the accuracy of the regression either remains stable or reduces slightly as the number of the added features increases. The visual inspection reveals that the increase in \( R^2 \) is relatively small (\( \ll 0.01 \)) when more input features are used. The NDVI regression rapidly converges, with typically 5 features, compared to the LAI and the BBCH regression. The highest number of features is required for an accurate estimation of BBCH stage. This is due to the fact that the first features are not very dominant in terms of the SI values, compared to the case of NDVI and LAI. In other words, the SI for the remaining features have almost the same values, around 0.2, meaning that they are equally important in the regression, whereas the SI of the rest of the features in the case of NDVI and LAI are quite close to zero. The backscattering ratio \( HH/HV \), (f9), has an impact on LAI estimation, however neither on NDVI nor on BBCH stage.

V. CONCLUSION

This study proposes a machine-learning based polynomial expansion to estimate crop biophysical variables from PolSAR data, and including an identification of the important PolSAR features through a GSA. The results show relevant and informative features, which may differ for each biophysical variable, provide a satisfactory prediction performance. HDMR-RR achieves an accurate prediction performance just based on a few of the most important features, hence a dimensionality reduction is achieved, since the rest of features do not contribute significantly to the estimation.

Results highlight that the co-polar backscattering ratio is the most important polarimetric feature for LAI, NDVI and BBCH estimation of barley. Additionally, the dominant scattering mechanism in time, i.e. the mean alpha angle plays a key
role in BBCH-scale estimation, but it can be ignored in the context of LAI and NDVI estimation. For LAI and NDVI, the features related to the volume, i.e., entropy (F5), as well as the ratio of $V/V'$ (F9) have a higher weight compared to the rest of the features.

It should be noted that precision of the parameters obtained by the HDMR-RR are expected to be relatively less accurate than those of parameters calculated by the classical HDMR, since it finds the parameters in the least square senses. Besides, to make more reliable analysis especially for the highly correlated datasets, the proposed approach needs to be modified in a way that takes into account the correlations between the features.

ACKNOWLEDGMENTS

The authors would like to thank the support of the Scientific Research Projects Coordination of Istanbul Technical University under Project MGA-2018-41152. This work was supported by the Spanish Ministry of Economy, Industry and Competitiveness (MINECO), the State Agency of Research (AEI) and the European Funds for Regional Development (FEDER) under Project TEC2017-85244-C2-1-P. Radarsat-2 Data and Products ©MacDonald, Dettwiler and Associates Ltd. (MDA, 2009) – All Rights Reserved. Radarsat is an official trademark of the Canadian Space Agency (CSA). All Radarsat-2 images have been provided by MDA and CSA in the framework of the ESA funded AgriSAR2009 campaign.

REFERENCES