Improving Utility of Low-Resolution Data using Statistical Approaches in Remote Sensing

Anuj Katiyal [1]
Krishnan Sundara Rajan [2]

International Institute of Information Technology
Hyderabad, India 500-032
[1] anuj.katiyal@research.iiit.ac.in, [2] rajan@iiit.ac.in

Abstract. With the increase in the multi-resolution data available from the various satellite sensors, there is an increasing need to come up with analysis techniques to handle and exploit the information that can be extracted from lower resolution (LR) data before acquiring higher resolution (HR) data. This paper presents a methodology to use statistical approaches to sub-group the LR classified data into high importance local regions (HILRs) and low importance local regions (LILRs) after filtering, for every class. The HILRs were shown to have more number of near pure-pixels as compared to the complete class regions, as verified by using the classified HR APLULC data, when a LR pixel was matched to the HR matrix using the Near Purity Measure (of 80). The HILRs were further shown to have higher stability, by showing reduced NDVI variation as compared to the complete class regions, using the HR AWIFS data. The method proposed works better for LR classes with limited intra-class heterogeneity and good inter-class separability. The proposed approach can help to reduce the processing done on HR resolution data based on the corresponding LR HILRs obtained for every class regions and further help in applications like pure-pixel matching, building HR-LR classification models and isolating pure pixels from the mixed/impure pixels in class regions.

Keywords: Multi-resolution, Multi-sensor, Multi-spectral, Sub-pixel classification, MODIS, AWIFS

1. Introduction

The last few decades has seen the growth of Earth Observation Systems (EOS) and many satellites have been launched, generally with more than one sensor for capturing image data at various resolutions. This available multi-resolution data holds great potential to support the national economy in the areas of agriculture, water resources, forestry and ecology, geology, marine fisheries, coastal management etc. and has been widely used in studies of environmental changes and impact analysis, natural resource management, ecosystem and land usage analysis. The spatial resolution of the data captured by various satellite sensors varies from 0.5m to 25000m (Chen, 2003). With the advancement in technology in the recent years, there has been an increase in the amount of data acquisition at an increased range of spatial resolutions for the same ground locations to capture more information. So, why should we really capture data at low resolution (LR) when data is available for the same spatial locations at higher resolution (HR)? The availability of data at increased spatial resolution has raised questions regarding the usability of data acquired at low resolution (LR).

Efficient analysis of the available multi-resolution data to improve the land use/cover mapping and linking thematic maps generated between finer resolution and coarser resolution has become a challenge (Foody, 2002). With the increment in the availability of HR data and LR data from the satellite sensors such as AWIFS (56m spatial resolution), MODIS (250m spatial resolution), LISS-III (23.5m spatial resolution), LISS-IV (5.8m spatial resolution) etc. there is an increasing need to develop efficient multi-resolution processing techniques. But, in studies like pure vs. mixed/impure pixel analysis, land use/land cover (LULC) analysis, the availability of HR satellite data for a particular period and region remains a challenge due to long sensor revisit times. Also, high cost of acquiring and processing HR imagery still makes most studies prefer LR satellite data available with frequent revisit times, in addition to its cost and computational advantages.

The various application areas in the field of remote sensing are affected by the scale variation and sensitivity of the available data (Marceau et al, 1994; Atkinson and Curran, 1999;

Chen and Stow, 2003). In the past years, many methods of image classification to improve the data accuracy of land use/land cover (LULC) have been developed, with the focus to improve the global class accuracies with limited applicability. It's a general trend to use the acquired HR data to report the improved classification accuracies for a particular interest area, but due to the heterogeneity of spectral-radiometric characteristics in the natural land cover as captured with the HR images, classification approaches using a particular (single HR) resolution data have not led to satisfactory results (Chen and Stow, 2003). Also, a generalized classification method cannot work for all spatial resolutions for the available remote sensing data (Marceau et al., 1994). Image classification outputs at LR (coarse) are often combined with the classified HR data (finer) to improve the results of the application, thus taking the advantages of multi-sensor, multi-scale and multi-temporal satellite imagery (Solberg et al., 1996; Li et al., 2000). However, compared with rapidly expanding multi resolution data sets, there is a need to develop more spatio-analytic methods and models (Quattrochi and Goodchild, 1997; Dungan J.L., 2001; Tate and Atkinson, 2001).

The objective of this research work was to develop a spatio-analytic method using LR sensor data, which mostly is readily available at frequent intervals and is generally free of cost, to come up with high importance local regions (HILRs) within classified class regions. The classified LR image regions can be sub-grouped at pixel level as pure pixels, mixed pixels and impure pixels. If we have HR classified data geo-registered with the LR classified base data, then pure pixels match only a particular class in corresponding HR data (or the base class), mixed pixels match multiple classes including the base class in HR data and impure pixels match multiple classes other than the base class in the corresponding HR data. We defined a term near-pure pixels (NPP) to refer to pure pixels (ideal case) having a particular (or base) class pixels greater than a threshold, in the corresponding classified HR data. The HILR regions are the estimated regions within the classified data having high percentage of near-pure pixels (NPP) and showing more stability (less NDVI variation) as compared to the complete class regions. HILRs can be useful in studies to isolate near-pure vs. mixed/impure pixels, as training samples for the classification of higher resolution (HR) imagery, in developing HR-LR hybrid classification models etc.

1.1 Study Area and Data

The study area used in this work corresponds to the region around Nagarjuna Sagar Dam, which lies in the state of Andhra Pradesh, in the south-eastern part of India (Figure 1). The experimental area considered corresponds to the co-ordinates 16°40' N, 78°17' E and 14°59' N, 80°16'E, and covers a total area of 36,646 km^2.

In order to perform the study, daily Terra/MODIS (at 250m spatial resolution) atmospherically corrected surface reflectance data (Figure 1) was acquired (product MOD09GQ) containing red (645 nm) and near-infrared bands (858 nm) for the dates 30 January 2006 and 1 January 2007. MODIS low resolution (LR) data had Sinusoidal projection with WGS-84 datum. For the same region and dates, AWIFS data (at 56m spatial resolution) having Lambert Conformal Conic projection with the WGS-84 Datum (Figure 1) was also acquired. The size of the HR AWIFS and LR MODIS data pairs, for the acquired dates, covering a portion of the south-eastern state of Andhra Pradesh, India, is 3319 × 3713 and 804 × 899 respectively.

Other than that, detailed documentation and LULC classified data with 18 classes for the state of Andhra Pradesh (APLULC), India was obtained from NRSC, Hyderabad for the year 2005-06, which was based on AWIFS imagery and extensive field visits. The classified HR APLULC data was further clipped to match and geo-reference the experimental region for the available LR MODIS and HR AWIFS data. The experiments were performed on the LR

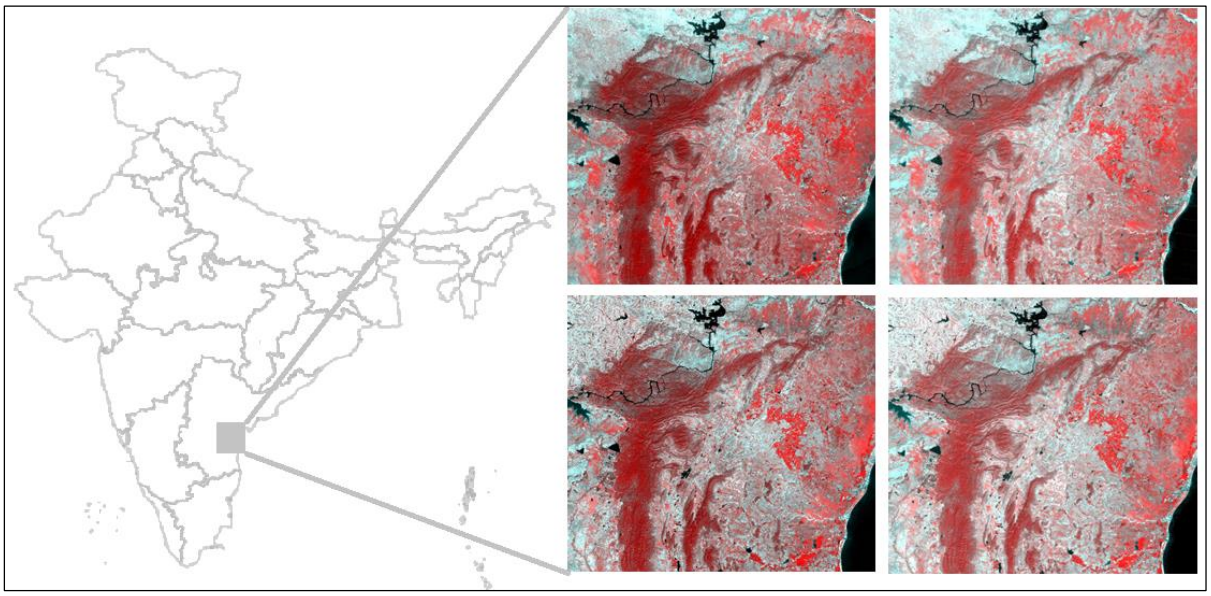

Figure 1. False Color Composite data showing the work area in SE India, (top left) AWIFS 1 January 2007, (top right) MODIS 1 January 2007, (bottom left) AWIFS 30 January 2006 and (bottom right) MODIS 30 January 2006.

MODIS data and validation was done using the HR classified APLULC data and NDVI values calculated for HR AWIFS data obtained for the regions (as shown in Figure 1).

2. Methodology

Unlike the general trend of acquiring HR data for a particular spatial location and improving the classification accuracy of the region, we tend to move our focus solely on gathering information from the available LR data (Gupta and Rajan, 2009), by detecting regions of higher confidence (HILRs) within each class of the classified LR data. By higher importance local regions (HILR), we refer to the regions within each class with comparatively less pixels as compared to the total class pixels, but increased stability, class accuracy and near-pure pixels (NPPs).

2.1 Data Pre-Processing

The pairs of HR and LR images for both the dates were converted to Lambert Conformal Conic Projection with WGS-84 Datum using the nearest neighbor approach. False color composite images were obtained by using the available bands for both the MODIS LR and AWIFS HR data for further processing and visualization. The HR APLULC data classes were combined into 5 broader class regions based on the hierarchy of the classes present in the data and the classes present in the experimental work region. Also, LR data being geo-registered with HR data, was pixel matched using nearest neighbor algorithm, i.e., after pixel matching, every pixel in LR data pixel corresponds to a matrix in HR data (as in Figure 2).

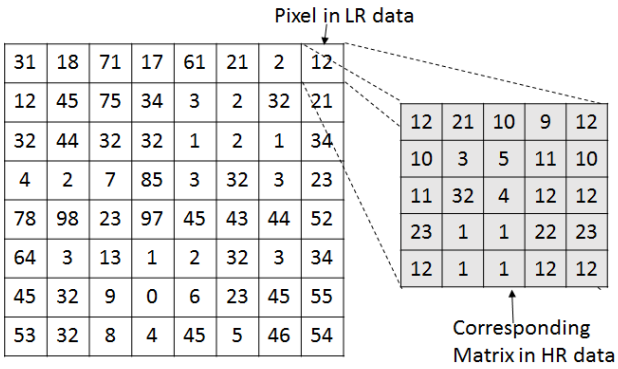

Figure 2. A random matrix showing correspondence between a LR data pixel to a HR data matrix

Further, the Normalized Differential Vegetation Index (NDVI), which is based on a ratio of the red and near-infrared bands was computed for the HR AWIFS data.

2.2 Filter Spectrally Classified Spatial Clusters

The pre-processed LR MODIS data available was first spectrally classified using unsupervised K-Means algorithm. The documentation available for HR classified APLULC data helped in deciding the parameter for K-Means algorithm, i.e., the number of classes; the change-threshold was kept at 5%. The number of classes were kept less than the total number of APLULC classes documented based on the classes present in our work-area and minimal hierarchy of classes shown by LR data as compared to HR data in multi-resolution images (Lin et al, 2003).

The resultant LR MODIS classes were further segmented into its connected components, which were called the spatial clusters (SC) of a class. The connected components obtained for every class were labelled with a unique SC number. The SCs obtained were filtered to ignore the ones with number of pixels below a threshold, which was experimentally observed based on a Near-Purity Measure (NPM) defined for every geo-registered pixel of LR MODIS to the corresponding matrix in HR classified APLULC data. Near-Purity Measure (NPM) of 'X' labels a pixel as near-pure pixel if (greater than) X% of the total corresponding HR matrix pixels belong to the same class. It was experimentally observed that increasing the threshold to ignore the SCs with less pixels, improved the percentage of near pure-pixels (NPP) (calculated using NPM of 80) to the total number of pixels in a class, for most LR MODIS classes. After filtering, representative SCs for every LR MODIS class were obtained.

2.3 High Importance Local Regions (HILR) within SCs

The algorithm used for obtaining HILRs within SCs was divided into two parts, the sampling from the SCs to obtain statistical measures and the labeling of SCs based on the measures calculated. Sampling algorithm used can be summarized as:

- For every SC in LR MODIS class, random sampling with replacement was done based on the following variables, the number of samples (NS) and the size of the samples (SS). Samples of size SS (greater than 100) were taken for NS number of times, for every SC.
- For every sample, statistical measures, namely, mean, median and mode were calculated. This resulted in NS statistical measures for every SC in a class.
- The mean pixel-value of the NS statistical measures (mean, median and mode taken one at a time) was taken as point of origin for the expansion of the local important regions within each SC.

The mean along with the standard deviation of the data pixels in the SC can be used to calculate the standard score (z-score). In statistics, the standard score (z-score) indicates by how many standard deviations an observation is away from the mean. The standard score (z) of a data pixel (x) is defined using mean of the population (μ) and standard deviation of the population (σ), as

$$z = \left(\frac{x - \mu}{\sigma}\right) \Rightarrow [x = (\mu + z \times \sigma)] \quad (1)$$

The result of the sampling algorithm was used as input to the iterative labeling algorithm, as described below:

- For every SC, the point of origin for the statistical measures was taken as the mean value and standard deviation for the complete SC was calculated. The z-score was calculated for every data value in the SC.
- For the algorithm, the z-value was incremented iteratively and the data values lying within the range of $\mu \pm z \times \sigma$ were labeled incrementally, with the label value increasing as the z-value increases.
- The lower the z-value, the lower the label, i.e., the lower is the distance of the data pixels from the mean. Using the above algorithm, the complete SC was divided into labels, i.e., different labeled SCs were obtained for different initial points of origin obtained using the sampling algorithm.

- Using an area threshold, the labels having sum of pixels less than it were the calculated high-importance local regions (HILRs). The labels for the data pixels above the area threshold were combined to form low importance local regions (LILR) within the respective SC.

For visualization purposes, the image was divided into green-regions (GR) and yellow-regions (YR) (shown in figure 3), as:

$$\forall_{(c=LR\ class)} Image_{(GR)}(c) = \forall_{(c=LR\ class)} \forall_{(sc=SC)}\ HILR(c, sc) \quad (2)$$

$$\forall_{(c=LR\ class)} Image_{(YR)}(c) = \forall_{(c=LR\ class)} \forall_{(sc=SC)}\ LILR(c, sc) \quad (3)$$

$$\forall_{(c=LR\ class)} Image_{(Combined)}(c) = \forall_{(c=LR\ class)}\ (Image_{(GR)}(c) \cup Image_{(YR)}(c)) \quad (4)$$

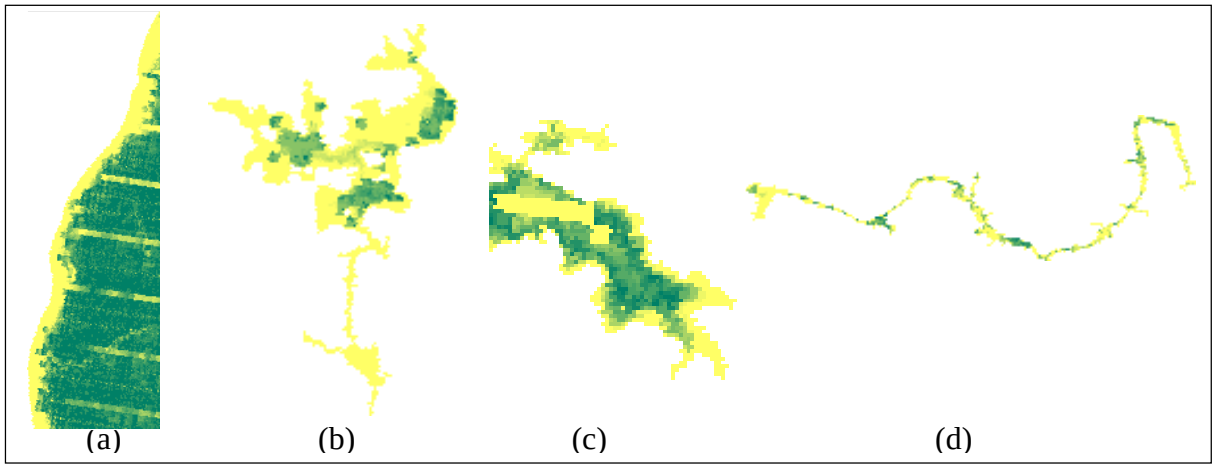

Figure 3. Four different Spatial Clusters for LR MODIS Class 1, divided into their yellow regions (YR) and green regions (GR) for 1 January 2007, using the mean of the SC sample modes as point of origin.

3. Results and Discussion

The verification of the green regions, i.e., the HILR for every SC within each LR MODIS class was done at the class level as well as the SC level. The verification at the class level was done by showing the increase in the percentage of Near Pure Pixels (NPP) in the green regions (Equation 2) as compared to the combined regions (Equation 4) for every class, calculated using the NPM (of 80). The SC level verification was done using the HR NDVI values calculated for the same region, by showing the decreased variation in the NDVI values for the HILR as compared to the complete SC region. The class regions in LR MODIS data represented water-bodies (class 1), forested regions (class 2), Rabi crop-lands (class 3), agricultural fallow lands (class 4) and others (class 5).

3.1 Class level verification using HR classified APLULC data

The first method uses HR APLULC classified and geo-registered data, to verify whether HILRs are good estimates of Near Pure Pixels (NPP) at a given NPM. The NPP, total pixels (TP) along with the percentage $\left({}^{NPP}/_{TP}\right) \times 100$ are calculated for the combined regions and the green regions for every class. Class regions were obtained using the mean of – (i) the sample means of the SC (SM1); (ii) sample medians of the SC (SM2); and (iii) the sample modes of the SC (SM3) as the points of origin calculated by the sampling algorithm. The results for the data pairs of 1 January 2007 having 5 LR MODIS classes, show an increase in the percentage (NPP to TP) from Combined Regions to Green Regions for all the experiments SM1 – Mean, SM2 – Median and SM3 – Mode for the classes 1 and 4 as shown in Table 1. Partial increment, i.e., increment in the percentage for one or more of the experiments was observed for classes 2 and 3 while no increment was shown for class 5. The classes with a high percentage (%) of NPP to TP (>= 60) in the combined regions, generally tend to show an increased percentage in the HILR obtained for each class in the conducted experiments. Classes with the

Table 1. NPP, TP and Percentage (NPP/TP) × 100 for the combine regions and the green regions obtained using mean of the SMs (mean, median and mode) as points of origin, for 1 January 2007 data.

					SM1 - Mean			SM2 - Median			SM3 - Mode		
		Combine			Green			Green			Green		
Class	Regions	NPP	TP	%	NPP	TP	%	NPP	TP	%	NPP	TP	%
1	Waterbodies	24307	26008	**93.46**	20006	20795	**96.21**	19386	20136	**96.28**	18624	18952	**98.27**
2	Forests	91180	112445	**81.09**	35263	43298	**81.44**	36176	44689	80.95	37172	45800	**81.16**
3	Crop Lands	11950	22009	**54.3**	4740	8795	53.89	4913	9098	54	4264	6717	**63.48**
4	Fallow Lands	42411	94246	**45**	16788	35174	**47.73**	16907	35709	**47.35**	13268	29059	**45.66**
5	Others	31435	82936	**37.9**	10734	28691	37.41	11442	30397	37.64	11043	29501	37.43

percentage in the range of 40-60% for combined regions, show partial increment in the percentage of green regions in the conducted experiments. Further, classes with less than 40 percent ratio of NPP to TP in combined regions itself, are bad candidate classes for the algorithm. More NPPs in a class refer to more homogeneity in the class, and in a way, decreased amount of mixed pixels. From Table 1, the analysis shows that our method works well for classes with more homogeneity and is able to detect HILRs within SC for such classes. The classes with higher percentage, namely 1, 2, 3 and 4 were selected for further validation using available HR AWIFS data and their combined class image (yellow and green regions) for experiments SM1-Mean, SM2-Median and SM3-Mode has been shown in Figure 4, Figure 5 and Figure 6 respectively.

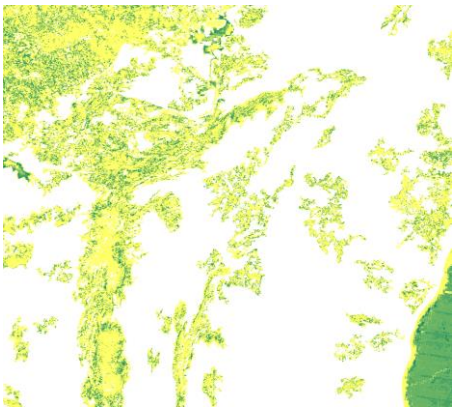

Figure 4. HILRs and LILRs for Classes 1, 2, 3 and 4 for SM1-Mean Experiment

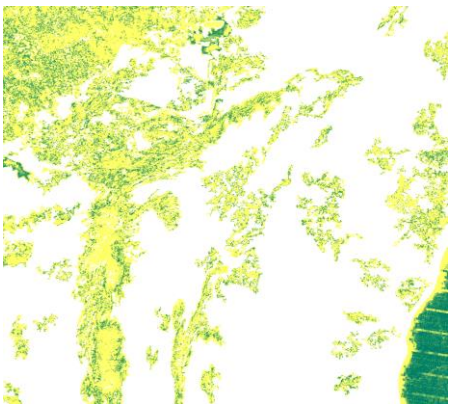

Figure 5. HILRs and LILRs for Classes 1, 2, 3 and 4 for SM2-Median Experiment

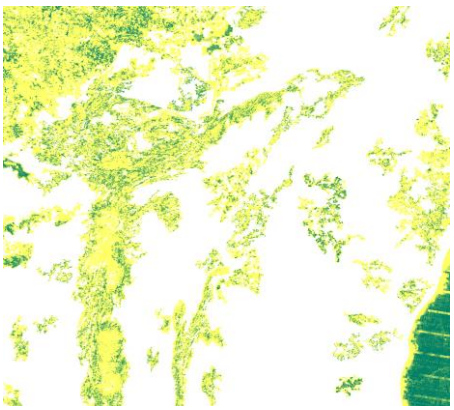

Figure 6. HILRs and LILRs for Classes 1, 2, 3 and 4 for SM3-Mode Experiment

3.2 Spatial Cluster (SC) level verification using HR AWIFS data

Further validation on the results was done at the level of every SC within every LR MODIS class, using the calculated HR AWIFS NDVI data obtained during the pre-processing phase. For the homogeneous classes having high NPP in the LR MODIS data, the green regions and the combined regions were matched with the corresponding HR NDVI regions. The range of observed NDVI values in the HR data matching the green regions and the combined regions for every SC of LR data was calculated. For the validation purpose, we calculated the mean and the standard deviation of the NDVI values calculated for green regions in every SC, and varied the z-score to get a range of NDVI values around the mean NDVI value. This approach helped in clipping the z-score range around the mean NDVI value and incrementally increasing the z-

score lead to an increase in the range of NDVI values considered (as shown in Equation 1) for every iteration. In each iteration, the HR AWIFS pixels matching the green and combined region, were counted for the range of the NDVI values. The iterations were stopped till majority (around 80%) of the green pixels belonged to the range of NDVI values. For all the classes having high NPP, namely classes 1, 2, 3 and 4, for an increment of the standard score (z-score) by 0.75 in every iteration, sample results for SCs in every class have been shown in Table 2 (experiment SM1-Mean for 1 January 2007).

Table 2. For Experiment SM1-Mean, the SC of LR MODIS classes with its mean and standard deviation through the iterations of the validation algorithm. The start range and the end range of NDVI values are given for every iteration and the result showing high percentage of matched HR green regions to total green regions as compared to matched HR combine regions to total combined region pixels has been highlighted.

Class, SC (μ,σ)	Iteration	Start_range	End_range	z-value	HR green match	HR total green	%	HR combine match	HR total combine	%
1,1 (-0.85, 0.26)	1	-1.05	-0.66	0.75	258949	291079	**88.96**	282000	332827	**84.73**
1,2 (-0.45,0.07)	1	-0.54	-0.37	0.75	23041	36578	**62.99**	37034	63752	**58.09**
1,2 (-0.45,0.07)	2	-0.62	-0.29	1.5	33296	36578	**91.03**	54771	63752	**85.91**
2,1(0.60,0.15)	1	0.49	0.71	0.75	176764	324386	**54.49**	361915	873790	**41.42**
2,1(0.60,0.15)	2	0.38	0.83	1.5	281258	324386	**86.70**	633199	873790	**72.47**
2,2(0.72,0.11)	1	0.61	0.84	0.75	28568	54670	**52.26**	61822	157749	**39.19**
2,2(0.72,0.11)	2	0.50	0.96	1.5	46454	54670	**84.97**	114017	157749	**72.28**
3,1(0.52,0.12)	1	0.43	0.61	0.75	35432	55814	**63.48**	74194	147262	**50.38**
3,1(0.52,0.12)	2	0.34	0.70	1.5	49106	55814	**87.98**	113822	147262	**77.29**
3,2(0.59,0.15)	1	0.48	0.71	0.75	9157	16448	**55.67**	19104	38729	**49.33**
3,2(0.59,0.15)	2	0.36	0.83	1.5	14832	16448	**90.18**	33917	38729	**87.58**
4,1(-0.05,0.07)	1	-0.11	0.00	0.75	222006	346996	**63.98**	507843	932820	**54.44**
4,1(-0.05,0.07)	2	-0.16	0.05	1.5	317197	346996	**91.41**	819464	932820	**87.85**
4,2(0.001, 0.09)	1	-0.07	0.07	0.75	18392	31073	**59.19**	46483	85375	**54.45**
4,2(0.001, 0.09)	2	-0.15	0.15	1.5	28166	31073	**90.64**	77228	85375	**90.46**

Similar validation was conducted for the classes having high NPP for the experiments SM2-Median and SM3-Mode for the available date pairs (1 January 2007 and 30 January 2006) and the results for all the SCs in the LR classes showed an increase in the percentage matching for the HR green regions as compared to the HR combine regions in the NDVI range for any iteration. Higher percentage for the HR green regions in every iteration showed their stability as compared to the HR combined regions. This was a further verification of the importance of the HILRs detected for SCs in LR MODIS data. These HILRs detected for classes with high homogeneity can be used for many applications like pure pixel matching, building LR-HR classification models, and isolating pure-pixels from impure/mixed pixels etc.

4. Conclusion

This research work was aimed at increasing the utility of available LR data because of its easy availability and advantages in term of cost of acquisition and processing. The usage of classified data has been generally limited by the global accuracy associated with every class, and we have shown that along with the global accuracy, processing done on the local regions within a class can lead to regions with better accuracy and higher utility for remote sensing data. It's a general approach to acquire higher resolution data for a particular place to improve the class accuracies, but we were able to show that for classes having limited intra-class heterogeneity and good inter-class separability, LR data can provide us with regions which need not require further processing in the corresponding HR data available.

In the current scenario, there is an increasing need to develop multi-resolution analysis techniques and come up with algorithms to extract as much information from the easily available LR data that can aid and supplement the analysis of the HR data. The future work lies in improving the sampling algorithm by deriving better points for origin, improving the labels on the SCs obtained and developing analysis techniques for various purposes based on the labels obtained. Also, handling of multi-resolution and multi-band data is a challenge.

Acknowledgements

We are grateful to NRSC, Hyderabad for providing the AWIFS spatial data.

References

Atkinson, P. M. Selecting the spatial resolution of airborne MSS imagery for small-scale agricultural mapping. **International Journal of Remote Sensing**, 18(9), pp.1903-1917, 1997.

Foody, G. M. Status of land cover classification accuracy assessment. **Remote Sensing of Environment 80**, pp.185-201, 2002.

Chen, D. and Stow, D. Strategies for integrating information from multiple spatial resolutions into land use/cover classification routines. **Photogrammetric Engineering & Remote Sensing**, 69(11):pp. 1279-1287, 2003.

Marceau, D.J., P.J. Howarth, D.J. Gratton. Remote sensing and the measurement of geographical entities in a forest environment 2: The optimal spatial resolution. **Remote Sensing of Environment**. 49, pp. 105-117, 1994.

Solberg, A.H.S.; Taxt, T.; Jain, A.K. A Markov random field model for classification of multisource satellite imagery. **IEEE Transactions on Geoscience and Remote Sensing**, 34(1):pp. 100-113, 1996.

Li, J., R.M. Gray, R.A. Olshen. Multi-resolution image classification by hierarchical modeling with two-dimensional hidden Markov models. **IEEE Transactions on Information Theory**, 46 (5):pp. 1826-1841, 2000.

Emerson, C. W., N. S-N. Lam and D. A. Quattrochi. Multi-scale fractal analysis of image texture and patterns. **Photogrammetric Engineering and Remote Sensing**, 65(1), pp. 51-62, 1999.

Dungan, J.L. Scaling up and scaling down: the relevance of the support effect on remote sensing of vegetation. In **Modeling Scale in Geographic Information Science**. Edited by N.J. Tate and P.M. Atkinson. John Wiley&Sons, Ltd., pp. 221-235, 2001.

Tate, N.J., and Atkinson, P. M. **Modeling scale in geographical information science**. John Wiley &Sons, 2001. 277p.

Schowengerdt, Robert A. **Remote Sensing Models and Methods for Image Processing**. Elsevier, 1997. 390p.

Gupta S. and Rajan K.S. **Exploring the Utility of Moderate Resolution Time Series Remotely Sensed Data for Land Use/Cover Classification**, 2009. 66 p. Dissertation (Masters in Computer Science Engineering) - International Institute of Information Technology, Hyderabad India. 2009.

Chen D. A Multi-Resolution Analysis and Classification framework for improving Land use/cover mapping from Earth Observation Data. In: The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences Volume XXXV Part B7, pp. 1187-1191, 2004. Available at: http://www.isprs.org/proceedings/XXXV/congress/comm7/papers/227.pdf Accessed on: 15 November 2012

Jessica Lin, Vlachos, M, Keogh, E., Gunopulos, D. Multi-resolution k-means clustering of time series and application to images. In Proceedings of the 4th SIGKDD Workshop on Multimedia Data Mining, in conjunction with SIGKDD 2003. Available at: http://www.cs.gmu.edu/~jessica/publications/ikmeans_mdm03.pdf Accessed on: 15 November 2012

Obtaining and Processing MODIS data. Available at: http://www.yale.edu/ceo/Documentation/MODIS_data.pdf Accessed on: 5 November 2012