STUDY AND IMPACT ANALYSIS OF DATA SHIFT IN DEEP LEARNING BASED ATMOSPHERIC CORRECTION

Maitrik Shah[1], Mehul S Raval[1], Srikrishnan Divakaran[2], Pragnesh Patel[1]

1. Ahmedabad University, Ahmedabad, Gujarat, India, 2. Krea University, Sri City, Andhra Pradesh, India

Ahmedabad University

I - INTRODUCTION

- Satellite images - affected by the presence of atmospheric elements.
- Atmospheric correction (AC) - process of removing the effect of the atmosphere and retrieving the correct surface reflectance (SR) values.
- DL based approaches can effectively emulate Physics based models

Image without AC (TOA)

Image with AC (BOA)

II - DATASET SHIFT

- The statistical properties of the satellite images evolve, causing dataset shift.
- Training data's statistical properties differ from testing data's.
- Ill-effects of dataset shift:
    - Poor generalization: The DL model fails to capture the test set relationships and will have performance degradation.
    - Bias introduction: Training and testing the DL model on different distributions may cause biased AC outcomes.
    - Unreliable model evaluation: The distribution mismatch makes assessing the model's true capabilities difficult.

III - DATASET DETAILS

- This paper uses Landsat 8 images from 2013 and 2020 captured during the winter season for six land cover(LC) types of the Indian subcontinent, with a cloud cover of < 15%.
- Each Landsat 8 scene with 30m spatial resolution covers an area of 185Km×180Km, translating into an image of size 7681×7531 pixels.
- Each image is divided into nonoverlapping patches of size 128x128, and the 400 most informative(based on entropy) are selected for the study.

IV - METHODOLOGY

We select an image based on location and year and extract patches from R, G, and B bands. Then we used three approaches to measure data set shift as follows:

DISTRIBUTIONAL LEVEL SHIFT MEASUREMENT

- Kolmogorov–Smirnov (KS) test is applied to cumulative distributive functions (CDFs) of training set joint histogram of (TOA,BOA) derived from patches of the year 2013 and testing set joint histogram of (TOA,BOA) derived from the patches of the year 2020.
- The KS test captures the overarching distributional changes with a measure of statistical significance. It measures changes in distributions' location, spread, or shape.

$$KS_{m,n,\lambda} = \max |F_{\lambda,J(TOA,BOA)}(x) - G_{\lambda,J(TOA,BOA)}(x)| \quad (1)$$

Land Covers	Blue Band		Green Band		Red Band	
	KS-value	p-value	KS-value	p-value	KS-value	p-value
Urban Land	0.22	1.4e-02	0.22	1.4e-02	0.22	1.5e-02
Crop Land	0.45	5.3e-05	0.45	5.4e-04	0.36	1.3e-03
Evergreen Forest	0.48	5.6e-07	0.47	6.1e-08	0.48	1.8e-07
Deciduous Forest	0.32	5.5e-03	0.31	6.0e-03	0.29	8.2e-03
Fallow Land	0.26	1.0e-02	0.27	5.9e-03	0.26	6.9e-03
Waste Land	0.35	1.8e-03	0.34	2.0e-03	0.33	1.2e-03

- Table shows the result of the KS test performed as per Equation 1.
- The p-value for all LC types across all bands is less than the a = 0.05.
- Also, all of them are smaller than the more conservative 1% a value.
- This strongly suggests that joint distribution of Test(TOA,BOA) is significantly different from Train(TOA,BOA).
- It is important to note that the KS test is influenced by the sample size; it is 128×128×400×3 per location in the present case. Its sensitivity to detect subtle differences increases as the sample size increases.
- Therefore, it is important to detect shift using another method, and we use the Wasserstein distance test to measure shift at a numerical level.

NUMERICAL LEVEL SHIFT MEASUREMENT

- Wasserstein distance between the difference of TOA, BOA patches of year 2013 and 2020 is calculated to measure the shift magnitude.
- The TOA−BOA provides the direct measure of numerical discrepancy evolving with time.
- It helps to understand the absolute differences in AC and captures drift more directly.
- It provides a quantifiable difference between the dissimilarity and independent of the distribution's shape and location.

$$WASS_{\lambda} = \int_{x \epsilon R} |F_{\lambda,TOA-BOA}(x) - G_{\lambda,TOA-BOA}(x)| \quad (2)$$

Land Covers	Blue Band	Green Band	Red Band
Urban Land	0.23	0.24	0.22
Crop Land	0.33	0.23	0.51
Evergreen Forest	0.56	0.55	0.50
Deciduous Forest	0.22	0.23	0.22
Fallow Land	0.40	0.31	0.079
Waste Land	0.25	0.22	0.14

- Table shows the distance calculated using Equation 2.
- A larger $WASS_{\lambda}$ suggests greater dissimilarity. It is seen from Table 3 that there are differences for all LC types across all bands.
- We can also observe that $WASS_{\lambda}$ is higher for the blue band and lower for the red band. The possible reason is that the blue band is most affected by atmospheric corruption.
- We also observe that distance is maximum for evergreen forests across all three bands and other locations have values between 0.22 to 0.44 for all LC types and across all bands.

DL MODEL LEVEL SHIFT MEASUREMENT

- We train a DL model using patches from 2020 which learns the relationship SR ≈ g(TOA) and test it on patches derived from 2013.
- We observe that RMSE is low when the model is tested on images of 2020. However, when tested on images from 2013, the RMSE is higher, indicating the drift in the dataset.

Land Covers	2020	2013	% Rise
Urban Land	0.045	0.050	10
Crop Land	0.045	0.043	-4
Evergreen Forest	0.038	0.047	23.6
Deciduous Forest	0.040	0.037	-7
Fallow Land	0.044	0.050	13.6
Waste Land	0.044	0.071	61

- It can be seen that for most LC types, the RMSE for 2013 is higher than the RMSE for 2020.
- In four cases, the % rise in RMSE for 2013 is > 10%.
- The RMSE marginally drops for cropland and deciduous forest, the reasons for which need to be investigated.

V - CONCLUSION

- We have implemented three different approaches to measure the data set shift and all three provide the compelling evidence of the presence of it.
- The approaches complement each other and strengthens the credibility and robustness of our findings.
- Though we have considered the inherent limitations of each test, this work needs to investigate further into the confounding factors and uncertainties with data and the tests employed.

ACKNOWLEDGEMENT

- Part of this work is carried out on the GPU purchased under Grant by the Department of Science and Technology-Gujarat Council of Science and Technology (DST-GUJCOST) under Grant GUJCOST/STI/2021-22/3858.
- The authors thank the School of Engineering and Applied Science, Ahmedabad University, and L D College of Engineering for providing the infrastructural support. The authors also thank SIPG group members at SAC ISRO, Ahmedabad, for their insightful comments

CONTACT: Maitrik Shah, maitrik.s@ahduni.edu.in