Conventional Urban Change Detection: The Impact of Spatial Resolution

Fanni Fiedrich[1]

Supervisor(s): Dr. Jan van Gemert[1]**, Prof. Dessislava Petrova-Antonova**[2]

[1]EEMCS, Delft University of Technology, The Netherlands
[2]GATE, Sofia University "St. Kliment Ohridski", Bulgaria

A Thesis Submitted to EEMCS Faculty Delft University of Technology,
In Partial Fulfilment of the Requirements
For the Bachelor of Computer Science and Engineering
June 23, 2024

Name of the student: Fanni Fiedrich
Final project course: CSE3000 Research Project
Thesis committee: Dr. Jan van Gemert, Prof. Dessislava Petrova-Antonova, Prof. Klaus Hildebrandt

Abstract

The detection of changes in an area over time using remotely sensed data such as images is referred to as change detection. It has a large range of applications. For example, changes in buildings can analysed for urban planning. Many conventional image processing and machine learning-based algorithms have been developed for the purpose of change detection. Conventional non-classification algorithms have advantages in their reduced computational cost. Remotely sensed images vary in their spatial resolution, which is the area a pixel covers on the Earth surface. This work aims to explore how the spatial resolution impacts conventional non-classification pixel-based techniques in the urban change detection context, to provide insight into their performance with regards to detecting urban-related change over different resolutions. A systematic experiment is conducted by considering the LEVIR-CD and OSCD test sets in their initial as well as multiple downsampled resolutions. The change detection algorithms Change Vector Analysis (CVA) and Iteratively Reweighted Multivariate Alteration Detection (IR-MAD) are applied to the data individually. For creating binary change labels on a pixel-level, the Otsu algorithm is applied. A set of performance metrics is calculated, and trends in the metrics values over the resolutions are analysed. The data shows some trends towards improved metric values for lower spatial resolutions. The degree of the trends varies and is dependent on the algorithm and dataset. Overall, further research is necessary to consider influencing factors such as the amount of pixels in images, to refine the processing steps, and to broaden the scope of the experiment.

1 Introduction

Remote sensing and Earth observation delivering satellite imagery or aerial photography has become of high importance in a large variety of fields [1, 2, 3]. Spatial data gathered over periods of time can be used to explore the development of regions, cities and countries [4]. Analysing changes in urban areas for urban planning can be helpful in adjusting to the needs of growing populations [3, 5]. Remote sensing in the urban context enables more frequent data collection than for example surveys and other administrative approaches [5], making it easier to detect changes consistently and over longer time periods. Other applications of remote sensing data are related but not limited to the preservation of ecological balance and climate change related research [3].

Change detection algorithms and models can be applied to a variety of spatial data like satellite images [4]. Depending on the precise algorithm, they aim at identifying change in a binary manner, or also the types of land cover changes that occurred. The central idea of change detection is visualised in Figure 1.

A great number of change detection algorithms and models using remote sensing data have been developed over time, and there are numerous literature reviews discussing the different approaches and solutions [1, 2, 3, 4, 6, 7]. In general they can be split into two groups: (1) traditional/conventional methods; and (2) Deep Learning (DL)/Neural Network-based methods. Some reviews specifically investigate DL [7] or Artificial Intelligence (AI)-based methods [1, 3], while others provide a general overview without focusing on a particular group of approaches [2, 4, 6].

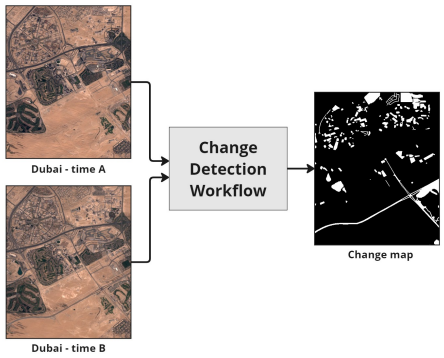

Figure 1: A simplified illustration of the central idea of change detection, using the OSCD dataset [8]. The left-hand side displays two images of Dubai taken at different points in time. The workflow includes pre-/post-processing and algorithm specific steps. The right-hand side shows the binary change map, white pixels signifying "change", and black pixels "no change".

DL techniques for remote sensing change detection are increasingly explored in recent research [7]. Among other aspects, DL approaches are more promising in terms of performance, and in their ability to represent semantic information [7]. However, DL based approaches are currently still facing some challenges. An important aspect to consider is the large amount of training data they can require [7]. Conventional algorithms that are based on algebraic operations or image transformations can directly provide a change map, and thus do not require any training data [6]. However, some aspects require manual consideration for conventional techniques. This includes choosing the algorithm itself, and the thresholds for change identification.

While there are efforts to solve the challenges DL faces in the future, aspects like training data and computational complexity are some of the trade-offs to be considered between DL and conventional change detection approaches [2, 7]. Therefore, it is of interest to not fully dismiss conventional techniques, and to spend efforts on exploring their capabilities and advances.

The spatial resolution of data is of great importance to remote sensing applications as different dataset resolutions are directly related for example to the active earth observation satellites, their orbits, and the current technological capabilities in terms of sensors [9, 10]. Comparing conventional techniques on their effectiveness dealing with varying spatial resolutions provides insight into the limitations and applications of each algorithm.

This work aims to answer the following main research question: *"How does spatial resolution impact non-classification conventional pixel-based techniques in the urban change detection context?"*. Spatial resolution refers to the resolution the data has on the Earth surface, indicating the information content of a pixel [9]. This work analyses the impact of the spatial resolution of data on the effectiveness of conventional algorithms. Experiments are conducted by ap-

plying different conventional algorithms on several datasets downsampled at varying resolutions.

The following specific research questions are defined to guide the study and analysis of results:

1. Is there a trend in falsely detecting change for unchanged pixels (*false positives*) when decreasing resolution?
2. Is there a trend in correctly detecting change for changed pixels (*true positives*) when decreasing resolution?
3. In general, how well do the algorithms identify the urban-related changes?

The paper is structured as follows. First, some additional background information is given to provide context on conventional methods. Next, the methodology of the experiment and the setup is laid out in detail with a focus on reproducability. Finally, the results are presented and discussed, as well as their limitations and possible future improvements.

2 Context of Conventional Change Detection

2.1 Algorithm Taxonomy

There are various taxonomies for change detection algorithms. Generally, conventional approaches can be divided into algebra-based, transformation-based and classification-based techniques [1, 6]. Algebra-based techniques involve performing algebraic operations on the images. For example, a naive approach is to subtract the images from each other, called image differencing [6]. Transformation-based methods involve a transformation step that reduces redundant or correlated information [1, 6]. Classification-based techniques either classify the data taken at one moment separate from the data at the other moment, and then compare the two resulting change maps, or merge the data, and classify the change directly [1].

Algorithms can also be divided on the level on which they consider data [4]. The algorithms presented in this paper detect change on a pixel-level. There are also techniques that detect changes on an object-level, first segmenting the image into different regions [4]. Each technique has different limitations and advantages.

2.2 Impact of Resolution

It seems there is a lack of systematic studies that review the impact of spatial resolution of data on conventional pixel-based algorithms. There exists one paper by ZhiYong et al. [11] which deals with different Change Vector Analysis (CVA)-based algorithms and found that their accuracy is dependent on the resolution, among other factors. The proposed methodology considers different datasets at varying spatial resolutions, instead of usage of a downsampled version of the same dataset.

3 Methodology

This section presents the methodology followed to conduct the current work aiming to investigate how robust a selection of conventional change detection algorithms are in handling different spatial resolutions of data. The methodology is visualised as a flowchart in Figure 2.

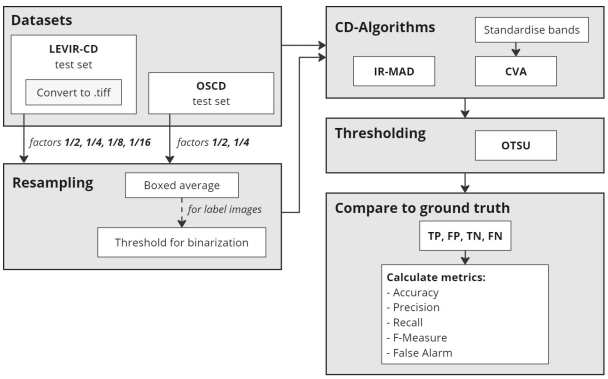

Figure 2: The methodology visualised in a flowchart to provide an overview at a glance. The datasets are downsampled, and used as input to the change detection algorithms. The output is compared to the ground truth, and performance metrics are calculated.

3.1 Data Collecting

The change detection datasets LEVIR-CD [12] and OSCD [8] are downsampled from their initial resolutions in multiple steps. Both datasets have a focus on urban related changes, for example building-related. They contain pairs of images taken at different times of the same area. For every image pair, a binary change map ("change"/"no change") is provided, which is labeled on a pixel-level.

The data with varying image resolutions and sizes is used to analyse how the algorithms perform depending on the resolution of the input. Table 1 provides an overview of the different spatial resolutions, scaling factors, and image sizes.

Instead of downsampling each dataset, an alternative method is to choose more datasets with varying resolutions to run the algorithms on [11]. However, every dataset has different properties that need to be taken into account. Aspects like types of areas that are imaged, and how the ground truth change maps are labelled, influence the results of the algorithms in addition to the resolution itself. To limit the confounding factors, only two datasets are chosen here.

The LEVIR-CD dataset is commonly used in recent research. Here, it is chosen because of its high resolution of 0.5 m/px [12]. This makes it possible to include more downsampling steps. However, a small image size of 64 by 64 px is already reached at a resolution of 8 m/px. In order to provide insights into algorithm performances at lower spatial resolutions, the OSCD dataset [8] is also considered. The combination of the two datasets allows for subsequent spatial resolutions ranging from 0.5m to 40m.

There is a notable difference in the amount of spectral information in the datasets. The LEVIR-CD dataset only covers RGB (red, green, and blue) color channels [12], while OSCD has data for a total of 13 spectral bands (later also referred to as image bands) [8]. The OSCD dataset contains data from the Sentinel-2 satellites with information from visible to infrared wavelengths. Additionally, the spatial resolution varies from 10 m/px to 60 m/px depending on the band. In the dataset, all data is initially upsampled to a resolution of 10 m/px. The RGB bands have an initial resolution of 10 m/px

without upsampling. To limit the differences of the datasets, the comparative analysis presented here selectively uses the RGB bands for the OSCD dataset.

Dataset	Scale	Resolution [m/px]	Sizes [px×px]
LEVIR-CD	1	0.5 m/px	1024×1024
LEVIR-CD	1/2	1 m/px	512×512
LEVIR-CD	1/4	2 m/px	256×256
LEVIR-CD	1/8	4 m/px	128×128
LEVIR-CD	1/16	8 m/px	64×64
OSCD	1	10 m/px	241×385, (...), 824×716
OSCD	1/2	20 m/px	120×192, (...), 412×358
OSCD	1/4	40 m/px	60×96, (...), 206×179

Table 1: The table provides a systematic overview of spatial resolutions and image sizes of the data (LEVIR-CD test set [12], OSCD test set [8]), easing interpretation of the results and ensuring replicability.

3.2 Running Change Detection Algorithms

Each version of the datasets is used as input for two conventional algorithms, namely Change Vector Analysis (CVA) [13], and Iteratively Reweighted Multivariate Alteration Detection (IR-MAD, also known as iMAD) [14]. CVA is an algebra-based technique [15], while IR-MAD is a transformation-based one [1]. Both algorithms can take any number of spectral bands as input.

3.2.1 Change Vector Analysis (CVA)

For CVA, the change vector between the two images is computed [13]. The magnitude and angle of the vector is used to classify a pixel as changed or unchanged. CVA is chosen since it is widely applied along various contexts, with different extensions of the algorithm existing [11].

Here, before calculating the change vectors, z-score standardisation [16] is applied to all image bands by scaling to unit variance and subtracting the mean, similar to what is done in an ESA software [17]. The intention is to limit effects of varying light conditions, and to have equal contribution from the different spectral bands[1]. Limitations to the standardisation approach will be discussed in section 6.

3.2.2 Iteratively Reweighted MAD (IR-MAD)

IR-MAD is an extension of the Multivariate Alteration Detection (MAD) algorithm [14, 18]. First, linear combinations of all bands per image are calculated, with coefficients determined by Canonical Correlation Analysis (CCA) [14, 19]. In maximising their correlation, CCA makes the resulting images as similar as possible to each other. The components of the results are then not ordered by wavelength (spectral bands) anymore, but by their correlation, simplifying the detection of changed pixels.

The increase in correlation of the images before taking their difference results in decreased variance of the difference image. IR-MAD extends the procedure of the MAD algorithm by then iteratively reweighting the observations on a pixel level. Observations with less or no changes will be assigned a higher weight in the process. The weights can then be used to determine whether a change occurred for a pixel.

IR-MAD is chosen as a transformation-based algorithm to be compared with CVA as an algebra-based technique. An advantage IR-MAD is that it is invariant to linear (affine) transformations [14]. Consequently, it is able to detect change in images even if they were taken in varying atmospheric conditions or with different sensor calibrations [19].

3.3 Thresholding and Evaluating

The output of the algorithms is not yet a binary change map. In the case of CVA for example, the result is the magnitude and angle/direction of change on a pixel level per band [11, 13]. By a manual fixed threshold or an (un)supervised thresholding algorithm, each pixel is assigned a binary "change" or "no change" label. The choice of the threshold has a high impact on the accuracy of the output [6].

The widely used Otsu thresholding algorithm is applied to the average value of the values per pixel [20]. It aims to determine the threshold at which the intensities of the classes ("change"/"no change") are optimally separated in terms of intensity variances. One advantage of the Otsu method and other unsupervised techniques is that they do not require manual inspection to determine a suitable threshold.

Other alternatives include approaches based on the Expectation–Maximisation algorithm, or based on Markov Random Fields [21]. In the context of this research, Otsu was chosen due to its usage by ZhiYong et al. [11] and for other conventional change detection techniques [22].

Once obtained, the binary change maps are compared to the corresponding ground truth labels of LEVIR-CD [12], and OSCD [8]. The algorithms are then analysed in their performance over the different image resolutions and datasets.

4 Experimental Setup and Results

This section presents the experimental setup and the results obtained following the proposed methodology in section 3. The setup is explained in subsection 4.1, to enable replicabilty for other studies. The results are presented in subsection 4.2.

4.1 Experimental Setup

Python is used for the first step of converting and scaling the datasets. The datasets are scaled using the OpenCV resize function[2] with the INTER_AREA interpolation algorithm as it is recommended for downscaling images. It calculates the average of a surrounding block for each pixel to get the downscaled values [23]. Since this can result in values between "change" and "no change" for the labeled change maps, the resulting images are thresholded for binarisation purposes. A

[1] ChatGPT was used for this argument. The answer was critically reflected upon. For more details and the exact prompts, see section 9.

[2] See the documentation, accessed 03.06.2024.

pixel with a value greater or equal to the average of pixel values corresponding to "change" and "no change" is categorized as "change", otherwise it is categorized as "no change".

For the change detection algorithms, a change detection toolbox [24] written in MatLab is used, and adapted to the use-case. This toolbox is chosen, since it contains functionality to: (1) load data; (2) run multiple change detection algorithms; and (3) run different thresholding algorithms. The code is overall adjusted to simplify the evaluation process, though the implementation for the CVA and IR-MAD algorithms themselves is left unmodified.

Two dataset loaders are added: (1) a LEVIR-CD dataset loader; and (2) a OSCD dataset loader which only considers the RGB bands. Only including the RGB bands makes the data more similar to LEVIR-CD dataset, and excludes undesired effects of having varying initial spatial resolutions between bands. For both datasets, the test sets (128 image pairs for LEVIR-CD, 10 for OSCD) are used, since the algorithms require no training or tuning of hyperparameters.

The MatLab code is adjusted to retrieve and store the following absolute pixel values in Excel files, instead of directly calculating metrics:

- True positives (TP): "Change" identified as "change"
- False positives (FP): "No change" identified as "change"
- True negatives (TN): "No change" identified as "no change"
- False negatives (FN): "Change" identified as "no change"

The usage of absolute pixel counts allows for easier follow-up evaluation and calculation. The results are extracted and evaluated in Python, mainly using Pandas[3].

The full code of the project is available on GitHub[4]. Functionality for the preparation of the datasets, as well as for the evaluation is mostly written in such a way that it can be adapted for future research.

4.2 Results

The metrics used for the evaluation process are listed and explained in Table 2. While the accuracy can give a general impression of the results, it does not take the balance of the ground truth labels into account [25]. Additionally, the F-score, as well as precision and recall are considered. The precision and recall are required for the interpretation of the F-score, since their role in the calculation of the harmonic mean is otherwise unknown. The false alarm (FA) error metric is included to analyse the fraction of pixels that are incorrectly identified as "change".

To ensure consistent image pair sets per resolution and per metric in the comparison between them, if the calculation of a metric results in a NaN value for some resolution, the corresponding image pair is fully discarded from the evaluation. Here, such values are encountered for the recall and for the F-score when the ground truth has no "change" pixels. In total, 16 LEVIR-CD image pairs are discarded, with a remainder of 112 pairs. More details and an extensive list of discarded files are given in section 9.

To illustrate what the output for different algorithms can look like, see Figure 3. For this example, it is evident that the algorithms tend to classify more pixels as "change" than in the ground truth. This pattern can be observed throughout the data, and in the metric evaluation.

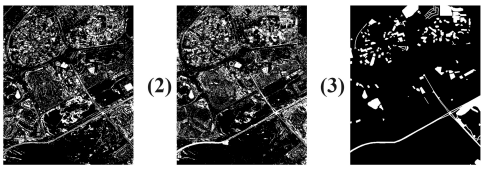

Figure 3: For illustration purposes, the resulting change maps for an example image pair of Dubai from the OSCD dataset at its initial resolution of 10 m/px are shown: (1) CVA results; (2) IR-MAD results; (3) ground truth. It is visible that the algorithms tend to detect additional pixels as changed, compared to the ground truth.

4.2.1 Correlations Between Spatial Resolution and Metrics

Correlation between spatial resolution and metrics is calculated per algorithm and shown in Figure 4. The correlations are presented not only for the combined datasets, but also split up for the LEVIR-CD dataset and the OSCD dataset. The split makes it easier to take the influence of dataset specific properties as well as their differing amounts of image pairs into account.

The notion of spatial resolution is important for the interpretation of the correlations. A lower spatial resolution means that the value in meters per pixel is higher, since more area is covered by a single pixel. Conversely, a higher spatial resolution has a lower numeric value.

Therefore, a positive correlation with spatial resolution for accuracy, precision, recall, and F-score indicates that the algorithms tend to perform better in these metrics with decreasing resolutions. Since FA is an error rate, a negative correlation indicates the same trend. Both of these aspects are indeed visible in the correlations, for all dataset combinations and metrics. The correlations vary from low to moderate.

There are differences in the amount of correlation when comparing groups of metrics. Recall tends to have less positive correlation than accuracy, precision and F-score. The exception is the data for CVA run on OSCD and LEVIR-CD individually. Specifically in the case of OSCD with CVA, recall exhibits a moderate positive correlation of 0.3, while accuracy, precision and F-score are in a low range between 0.035 and 0.049.

The amount of correlation also varies per algorithm/-dataset combination. Generally speaking, the correlations are stronger for the combined datasets. Again, there are some exceptions to this pattern, for example the recall value for applying CVA to only OSCD.

4.2.2 Mean Metric Values over Spatial Resolution

The mean value of each metric is plotted over the spatial resolutions in Figure 5. Up to 8 m/px the values correspond

[3] See the webpage, accessed 20.06.2024.

[4] See https://github.com/fa-fie/bsc-research-project.

Metric	Formula	Explanation
Accuracy [25]	$\frac{TP+TN}{TP+TN+FP+FN}$	Portion correctly identified pixels
Precision [25]	$\frac{TP}{TP+FP}$	Portion of pixels identified as "change" that were correct
Recall [25]	$\frac{TP}{TP+FN}$	Portion of "change" correctly identified as "change"
F-score [25]	$2 \cdot \frac{precision \cdot recall}{precision + recall}$	Harmonic mean of precision and recall
False Alarm (FA) [11]	$\frac{FP}{FP+TN}$	Portion incorrectly identified as "change"

Table 2: Short explanation and formula for each metric used in the evaluation process for a summarising overview.

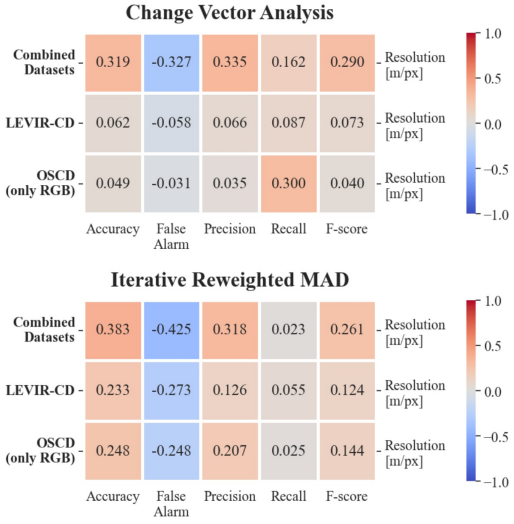

Figure 4: Matrices showing the correlations between metrics (accuracy, false alarm, precision, recall, F-score), and spatial resolution [m/px] for both algorithms (top: CVA, bottom: IR-MAD) to analyse the connection between resolutions and metrics. Each matrix contains data from both test sets (top), data from only the LEVIR-CD test set (middle) [12], and data from only OSCD test set using the RGB spectral bands (bottom) [8].

to LEVIR-CD, and from 10 m/px onwards to OSCD. The mean plots are used to provide an averaged indication on the amount of change that occurred, if any. In contrast to the correlations, they also show the progression over resolutions.

An important observation is that there is a difference in the values of most metrics when transitioning from LEVIR-CD to OSCD. This indicates that the dataset properties influence the results of the algorithms. The difference between datasets is always smaller for IR-MAD, compared to CVA.

Overall, the algorithms tend to perform somewhat similarly, recall being the exception. For recall, IR-MAD shows a recall value of approximately 0.18 higher than CVA in the LEVIR-CD resolutions. This indicates IR-MAD performs better at correctly identifying the changed pixels in LEVIR-CD image pairs. For the OSCD dataset sampled at 20 m/px, IR-MAD performs slightly worse than CVA in terms of recall.

The values for the F-score and precision are overall low. This indicates that the algorithms do not perform well at isolating the changed areas. Pixels that are unchanged in the ground truth tend to be identified as changed.

The accuracy has mean values spanning between approximately 0.7 and 0.9. This indicates that even though the algorithms struggle to isolate the changed area from the unchanged area, the overall portion of correctly identified pixels is relatively high.

For the majority of metrics there is a slight, though not consistent, upward or downward trend over the spatial resolutions. The trends are generally more apparent for IR-MAD than for CVA. For the F-score, precision, and accuracy there is a slight upward trend. For FA, there is a slight downward trend. Both of these aspects indicate a slightly improved mean performance for lower spatial resolutions. Recall shows a different pattern, but this was elaborated upon previously.

4.2.3 Metric Change Direction per Image

Figure 6 shows the percentages of image pairs for which each metric showed an in-/decrease, or remained unchanged, in comparison to the metric value of the same image pair at the *previous* resolution. There are no percentages shown for 0.5 m/px and 10 m/px, since they are the first resolutions of either dataset. The plots provide insight into the progression of the metric values on a more isolated basis, as they are evaluated individually per image, and not as a mean.

However, since the values themselves are not plotted, the precise metric differences cannot be analysed. For this reason, the discussion is held brief and the plots are only restrictively taken into account in the evaluation.

For the most part, accuracy, recall, precision and F-score tend more often to show increase over the resolutions on an image pair basis. This can be seen in the "increase" fraction being usually above 50%. Exceptions are 1 m/px and 20 m/px for recall (IR-MAD), 8 m/px and 40 m/px for precision (CVA), and 40 m/px for F-Measure (CVA). The false alarm predominantly decreases with "decrease" fraction being above 50%. The percentages show a difference in trends on a dataset basis.

5 Discussion

The research question will be discussed by investigating the hypothesis that the pixel-based conventional techniques utilised in the study will have a tendency to perform better

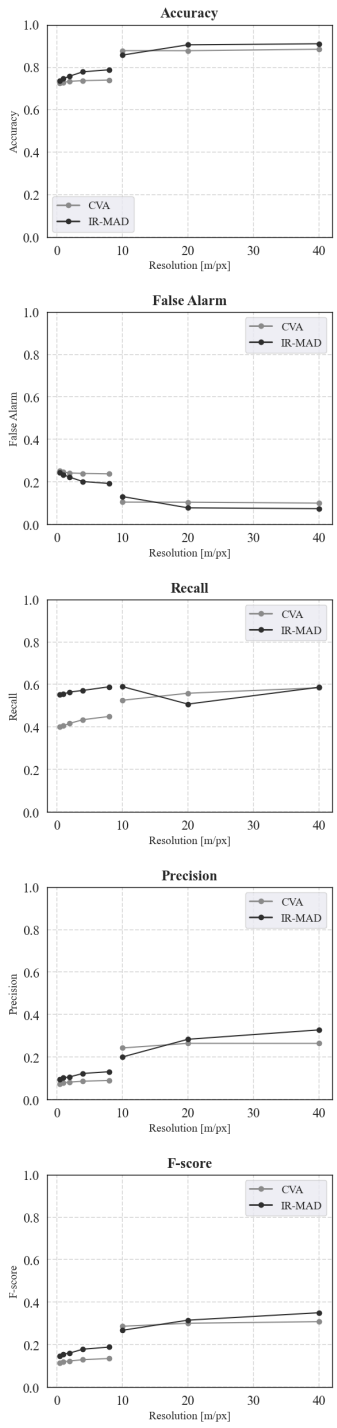

Figure 5: Mean accuracy, false alarm, precision, recall and F-score values for CVA (light grey) and IR-MAD (dark grey) over different spatial resolutions to provide an averaged indication on the amount of change that occurred, if any. From 0.5 m/px to 8 m/px, the values correspond to the LEVIR-CD test set [12], and from 10 m/px onwards to the OSCD test set (only RGB bands) [8].

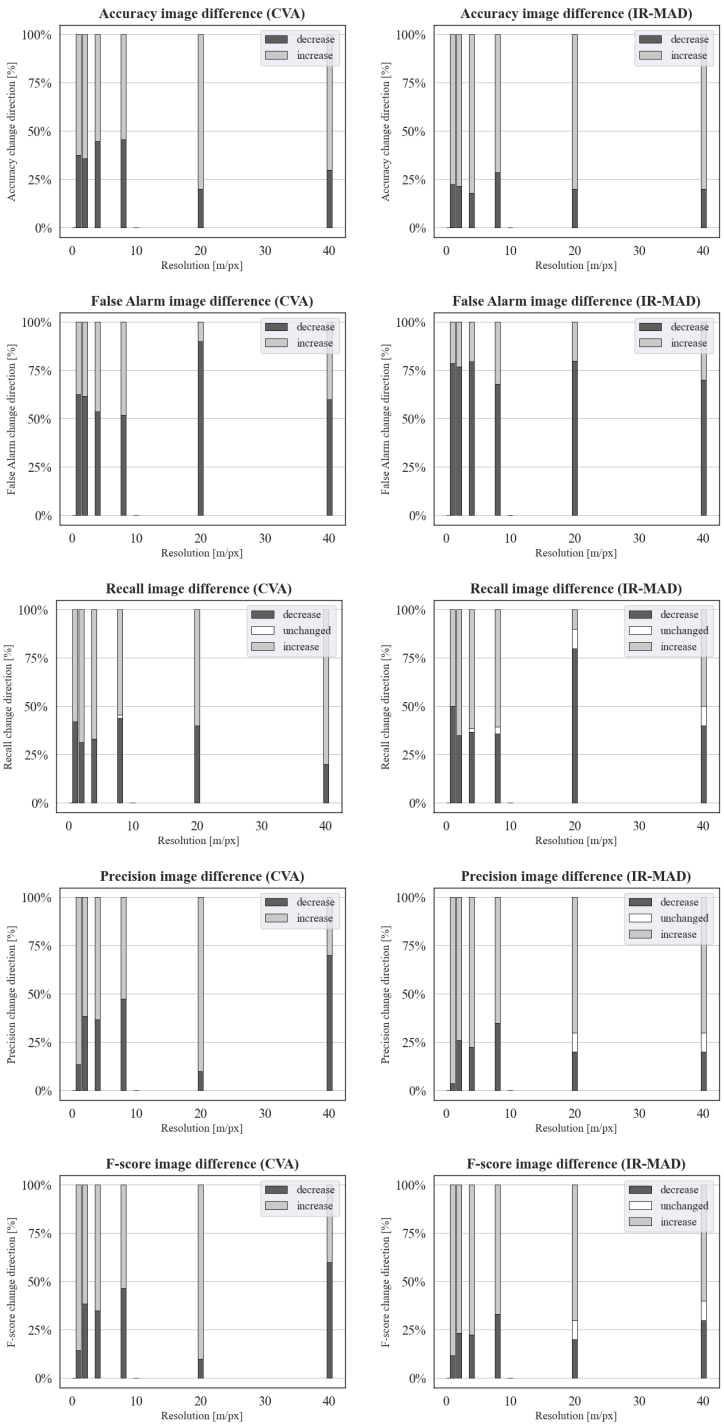

Figure 6: The proportion of image pairs for which each metric showed an increase (light grey), decrease (dark grey), or remained unchanged (white), in comparison to the metric value at the image pairs previous resolution. CVA is on the left, and IR-MAD on the right. The metrics are accuracy, false alarm, precision, recall and F-score. The plots provide insight into the general progression of metric values on a more isolated, image pair, basis.

for lower spatial resolutions in terms of the applied metrics. Different aspects can influence the performance.

For one, pixel-based methods are more susceptible to classify noise as change [2]. They face the problem of "salt-and-pepper noise" [3]. This refers to only partial detection of change for a changed object due to spectral heterogeneity. Another negative influencing factor for CVA and IR-MAD is that they do not consider the spatial context of a pixel. The downsampling process here might average noise out. Therefore, the impact of noise could be higher for the non-downsampled datasets.

However, an important aspect is that in lower resolutions, each pixel covers more area on the ground. This introduces the problem of "mixed pixels" in which one pixel contains mixed information belonging to different area types or objects [26]. Mixed pixels could cause the algorithms to be worse at identifying the corresponding changed areas. Therefore, the performance could also be negatively impacted for downsampled datasets.

5.1 Trend in False Change Detections

The false detection of "change" for "no change" pixels in the ground truth is related to the FP rate. The algorithms themselves do not use urban-related information. However, the datasets only label urban-related changes. This discrepancy between datasets and algorithms, and the false labelling of noise as change, as well as other factors, can cause the algorithms to detect "no change" pixels as "change".

The aim is to investigate whether there is a trend in such detections using FA and precision. The FA metric indicates the portion of "no change" pixels that were incorrectly identified as "change". The precision provides the portion of "change" detections that were correct.

FA slightly decreases over resolutions, but it is more evident for IR-MAD than for CVA. In the case of CVA, there is only a noticeable difference between the datasets (correlation of -0.327 for combined datasets), and not within the datasets (correlation of -0.058 to -0.031). For IR-MAD on the other hand, the correlation is generally moderate. It ranges from around -0.25 for individual datasets to -0.425 when combined. The tendency for IR-MAD to show a decrease in FA is also visible in the percentage plot with the decrease percentage fluctuating around 75%.

The precision tends to increase over resolutions, as seen in the positive correlations. Again, for CVA, there is only a difference between the two datasets rather than a noticeable increase within datasets. Further, a large fraction ($\approx$ 70%) of images shows a decrease in precision from 20 m/px to 40 m/px for CVA. IR-MAD shows a more visible trend with correlations of 0.126 (LEVIR-CD), 0.207 (OSCD) and 0.318 (combined).

Therefore, for these datasets and setup, IR-MAD tended to be more precise in the "change" predictions for lower resolutions. From the means it is visible that the difference in metric values is moderate to low. For CVA, the conclusions are quite limited. The only conclusion that can be drawn is that it performed better on the OSCD dataset.

5.2 Trend in Correct Change Detections

The portion of "change" pixels correctly identified as "change" or, conversely, those incorrectly identified as "no change" will be studied with respect to the recall metric. Here, the portion of additional incorrectly identified change is not considered. Therefore, precision and FA are not relevant. Analysing how well the algorithms identify urban-related changes can also be done by investigating recall, since the datasets both only label urban-related changes.

Overall, the correlation values for the recall are mostly low. For IR-MAD, there is a slight dip in performance for the OSCD dataset sampled at 20 m/px. This can also be seen in the fact that the majority of images showed a decrease rather than an increase in recall ($\approx$ 75%). With recall correlations of 0.055 (LEVIR-CD), 0.025 (OSCD) and 0.023 (combined) no trend can be found for IR-MAD. One interesting observation is that the mean recall values are close together at the transition from the LEVIR-CD dataset to the OSCD dataset.

For CVA on the other hand, a moderate correlation of 0.3 is found for the OSCD dataset. Even though a slight trend in the means can be seen, the correlation for the LEVIR-CD image pairs is low (0.087). For the combined datasets, it is 0.162. This indicates that here, CVA tends to identify more of the "change" pixels as "change" when decreasing the resolution in OSCD. It also shows that the impact depends on the dataset.

An important observation is that IR-MAD outperforms CVA in the mean recall values for the LEVIR-CD dataset. This could be because the pre-processing for CVA is limited to a simple standardisation of band values. More limitations of this aspect will be discussed in section 6. However, due to the restrictions of the project, no specific reason for the behavior could be identified with certainty.

5.3 Trend in Overall Performance

The overall performance of the algorithms on the datasets with the presented setups is low. This is evident from the low precision values, and in the mean recall values. The low precision can be due to the discrepancy between datasets and algorithms in whether urban-related information is considered. However, the recall value is also not high. Therefore, they are only able to partially identify the changes in general.

The mean accuracy ranges from approximately 0.7 to 0.9. Therefore, the percentage of correctly identified pixels is relatively high. However, it does not take the balance of labels into account. Both datasets have mostly "no change" pixels.

In reference to the initial research question and the investigated hypothesis, the main conclusions are as follows. The algorithms show a general tendency towards improved metrics when decreasing the resolution, to varying degrees. Within datasets, the absolute mean differences are low. The varying properties of the datasets have a clear influence on the metric values. Overall, the causes of the trends in the data need to be investigated more thoroughly, since various factors can influence them.

For CVA, there are no clear overarching trends within datasets that take all metrics into account, with noticeable differences between datasets. For IR-MAD however, there are

more consistent trends showing improvement for lower resolutions. The exception is the recall, so the correct detection of "change" pixels.

6 Limitations and Future Work

Overall, the main limitation is the rather small size of the experiment. In the following, the limitations to the presented research and methodology along with avenues for future work will be presented.

One central limitation to the experimental setup is that it should be extended by more advanced pre-processing steps. As mentioned in section 3, CVA is applied after a rather simple standardisation step. This choice was made at the beginning of the project, and an adjustment is not possible because of the limiting time frame. Pre-processing usually involves more complex radiometric and atmospheric normalisation or correction procedures to account for varying factors between the images (e.g., illumination, atmospheric conditions) [1, 13, 15]. Further, standardisation could influence the results in unforeseen ways, if the data is not normally distributed [16]. It is expected that the results for CVA could be improved and influenced by applying more advanced preprocessing steps. For IR-MAD the normalisation is not necessary, since the algorithm is invariant to relevant transformations.

Moreover, the size in pixels of the input images is not taken into account during the analysis. The reduced image sizes after downsampling are expected to have an impact on the results in addition to the resolution itself. To estimate the impact, a secondary experiment would need to be conducted. For example, this could be done by cutting the images into parts, taking the sections as input for the algorithms, and calculating the metrics by "re-assembling" the images. Here, this was also not possible due to the restricting time frame.

Another factor for the interpretation is the influence of label balance and amount of "change"/"no change" pixels. For example, if there are only two "change" pixels, one of which is correctly detected, the recall would be 0.5, though the information gained from that value would be limited.

As shown in the results, the algorithms themselves do not perform well on the datasets. They tend to detect more pixels as changed than the actual changed areas in the ground truth. In addition to the previously mentioned pre-processing limitations, this can partly be because of the simplicity of the algorithms themselves. The generalisability of the conclusions presented here is also limited since other algorithms might show different behavior.

Therefore, it is of interest to replicate the workflow with more complex algorithms in order to compare the results to those presented in this work and to be able to draw more nuanced conclusions. Apart from other algebra or transformation-based algorithms, or adding conventional classification-based algorithms further possibilities include:

- **Object-level algorithms**: These are less susceptible to the negative effects of "salt-and-pepper noise" [3]. However, they face different problems than pixel-based methods. Therefore, it would be of interest to investigate how they are influenced by a downsampling process.
- **Pixel-based methods with spatial information**: For example, an algorithm is developed by Kondmann et al. [22] that takes the context of a pixel into account. Since the downsampling influences the spatial context as well, it would be interesting to analyse how well it performs for varying spatial resolutions.
- **Algorithms with information based on indices**: There exist calculations on spectral bands that provide semantic information, referred to as indices. Various built-up indices are presented and discussed in a paper by Valdiviezo-N et al. [27]. Algorithms using semantic information could improve the performance of the simple CVA and IR-MAD methods presented here.

As outlined in previous sections, the usage of two different datasets limits the conclusions that can be drawn from the combined data due to varying dataset properties. Additionally, the factor of only including two datasets with non-overlapping resolutions influences the generalisability of the results. Due to the influencing factors of dataset-specific properties, the same setup with other datasets could show different patterns.

7 Conclusions

The research question "How does spatial resolution impact non-classification conventional pixel-based techniques in the urban change detection context?" is investigated by conducting an experiment involving the initial and multiple downsampled resolutions of the LEVIR-CD dataset [12] as well as the OSCD dataset [8]. The initial hypothesis is that the CVA and IR-MAD algorithms would have a tendency to perform better for lower spatial resolutions in terms of accuracy, false alarm, recall, precision, and F-score.

Varying degrees of trends towards improved metric values for lower resolutions are found. The metrics show a difference between the LEVIR-CD dataset and OSCD dataset, indicating influence of dataset-specific properties. Within datasets, the absolute differences in the averages of metrics are low. The conclusions differ for CVA and for IR-MAD. In the case of CVA, the correlation between metrics and spatial resolution is low per individual dataset. A clear increase in metric-specific performance is largely only observed when transitioning from LEVIR-CD resolutions to OSCD resolutions. Compared to CVA, the correlations within datasets for IR-MAD are mostly higher. The recall is an exception, showing only low, though positive, correlations for IR-MAD.

A central limitation of the research is the small size of the experiment, resulting in limited knowledge on the causes for the observed trends. Further research is deemed necessary to investigate influencing factors such as image sizes. Additionally, the experiment should be extended by pre-processing steps and including different datasets or types of change detection algorithms.

8 Responsible Research

Reproducibility is a central aspect of responsible research. The primary reason for the importance of reproducible research is that it makes the presented results verifiable. A

reproducible methodology also enables other researchers to replicate the workflow, and to build on the work more easily. Ethical aspects of the work will be considered in the following section as well.

8.1 Datasets

Both datasets can be openly accessed at the time of writing (that being June of 2024). The LEVIR-CD dataset [12] is available at https://chenhao.in/LEVIR/. The dataset consists of Google Earth image pairs, and the changes are labeled, as well as double-checked, by experts. The paper presenting LEVIR-CD has been cited more than 700 times, and the dataset is an often-used benchmark in the field. These aspects highlight the credibility of the dataset.

The OSCD dataset [8] is accessible on the IEEE DataPort [28]. It is published under a Creative Commons license. The image pairs are obtained from the Sentinel-2 satellites. Such data can be openly accessed. Similarly to LEVIR-CD, the ground truth labels are annotated manually, and only consider urban related changes.

It is safe to assume that both datasets are composed of ethically obtained image pairs. Both LEVIR-CD and OSCD use data which is accessible to the public, and usable for academic purposes. Generally speaking, a large part of raw and processed remote sensing data is openly accessible.

8.2 Methodology and Code

Generally, the methodology is explained in sufficient theoretical and practical detail such that the experiment can be replicated and built upon. The change detection toolbox [24] written in MatLab is published under the "Anti 996" license[5] as well as partially under the MIT license[6]. Both licenses are made to comply with open source standards, with the "Anti 996" license adding restrictions on the adherence to labor laws. The toolbox repository is openly accessible on GitHub at the time of writing (June of 2024).

To make reproducibility possible, the full code of the study presented in this paper is made openly accessible at: https://github.com/fa-fie/bsc-research-project.

Unfortunately, as with any type of code, no guarantees can be made about the code being bug-free. Specifically, detecting bugs in the change detection toolbox would be difficult. The reasons are it being external code, as well as the limited knowledge of algorithms and of the programming language MatLab at the start of the project. However, no influencing bugs are known.

8.3 Ethical Considerations

Since the algorithms themselves show only a low precision, the main ethical implications of the experiment are in its limitations. These are discussed in section 6. Due to the limited generalisability, the experiment should be extended and replication studies should be performed.

Moreover, there are some general ethical implications specific to Earth observation and remote sensing data, which are used as input to the algorithms. A recent recommendation paper from December 2023 by the Climate and Societal Benefits Subcommittee of the National Space Council Users' Advisory Group in the United States is taken as reference [29]. It discusses aspects specific to space data ethics, and argues why the subcommittee believes the development of a framework for such ethics is necessary.

A central concern is that remotely sensed data is usually "dual-use" [29], which is also the case for when change detection is applied to it. This means that it can inherently be used both for good-intentioned as well as ill-intentioned purposes. In the context of change detection this could for example be (1) gauging the impact of a natural catastrophe to find which regions need the most help; or (2) tracking the positions of potentially vulnerable groups.

Here, it is therefore argued that the ethical implications of remote sensing data applications need to be analysed on a case-by-case basis, since they can vary largely between scenarios. The application of change detection algorithms to data should be considered with a perspective on how it affects people in the real world outside of the theoretical context of the study.

[5] See https://github.com/kattgu7/Anti-996-License, accessed 03.06.2024.

[6] See https://opensource.org/license/mit, accessed 03.06.2024.

9 Acknowledgements

I would like to thank Prof. Petrova-Antonova for her guidance and Prof. van Gemert for his feedback throughout the project. Also, special thanks to Kayleigh Jones for insightful discussions, as well as feedback on the paper.

References

[1] W. Shi, M. Zhang, R. Zhang, S. Chen, and Z. Zhan, "Change Detection Based on Artificial Intelligence: State-of-the-Art and Challenges," *Remote Sensing*, vol. 12, no. 10, p. 1688, Jan. 2020.

[2] B. Farooq and A. Manocha, "Satellite-based change detection in multi-objective scenarios: A comprehensive review," *Remote Sensing Applications: Society and Environment*, vol. 34, p. 101168, Apr. 2024.

[3] Z. Gu and M. Zeng, "The Use of Artificial Intelligence and Satellite Remote Sensing in Land Cover Change Detection: Review and Perspectives," *Sustainability*, vol. 16, no. 1, p. 274, Jan. 2024.

[4] G. Cheng, Y. Huang, X. Li, S. Lyu, Z. Xu, Q. Zhao, and S. Xiang, "Change Detection Methods for Remote Sensing in the Last Decade: A Comprehensive Review," May 2023.

[5] K. J. Dueker and F. E. Horton, "Urban-change detection systems: Remote-sensing inputs," *Photogrammetria*, vol. 28, no. 3, pp. 89–106, Sep. 1972.

[6] Y. Afaq and A. Manocha, "Analysis on change detection techniques for remote sensing applications: A review," *Ecological Informatics*, vol. 63, p. 101310, Jul. 2021.

[7] T. Bai, L. Wang, D. Yin, K. Sun, Y. Chen, W. Li, and D. Li, "Deep learning for change detection in remote sensing: A review," *Geo-spatial Information Science*, vol. 26, no. 3, pp. 262–288, Jul. 2023.

[8] R. C. Daudt, B. Le Saux, A. Boulch, and Y. Gousseau, "Urban Change Detection for Multispectral Earth Observation Using Convolutional Neural Networks," in *IGARSS 2018 - 2018 IEEE International Geoscience and Remote Sensing Symposium*. Valencia, Spain: IEEE, Jul. 2018, pp. 2115–2118.

[9] J. Zhang and J. Li, "Chapter 11 - Spacecraft," in *Spatial Cognitive Engine Technology*, J. Zhang and J. Li, Eds. Academic Press, Jan. 2023, p. 142.

[10] A. Asokan and J. Anitha, "Change detection techniques for remote sensing applications: A survey," *Earth Science Informatics*, vol. 12, no. 2, pp. 143–160, Jun. 2019.

[11] L. ZhiYong, F. Wang, L. Xie, W. Sun, N. Falco, J. A. Benediktsson, and Z. You, "Diagnostic Analysis on Change Vector Analysis Methods for LCCD Using Remote Sensing Images," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 14, pp. 10 199–10 212, 2021.

[12] H. Chen and Z. Shi, "A Spatial-Temporal Attention-Based Method and a New Dataset for Remote Sensing Image Change Detection," *Remote Sensing*, vol. 12, no. 10, p. 1662, May 2020.

[13] W. Malila, "Change Vector Analysis: An Approach for Detecting Forest Changes with Landsat," *LARS Symposia*, Jan. 1980.

[14] A. A. Nielsen, "The Regularized Iteratively Reweighted MAD Method for Change Detection in Multi- and Hyperspectral Data," *IEEE Transactions on Image Processing*, vol. 16, no. 2, pp. 463–478, Feb. 2007.

[15] D. Lu, P. Mausel, E. Brondízio, and E. Moran, "Change Detection Techniques," *International Journal of Remote Sensing*, vol. 25, Jan. 2004.

[16] DATAtab Team, "Z-Score: Definition, formula, calculation & interpretation," 2024, accessed : 22 June 2024. [Online]. Available: https://datatab.net/tutorial/z-score

[17] European Space Agency (ESA), "CVA service specifications," accessed : 22 June 2024. [Online]. Available: https://docs.disasterscharter.org/services/cva/service-specs/

[18] A. A. Nielsen, K. Conradsen, and J. J. Simpson, "Multivariate Alteration Detection (MAD) and MAF Postprocessing in Multispectral, Bitemporal Image Data: New Approaches to Change Detection Studies," *Remote Sensing of Environment*, vol. 64, no. 1, pp. 1–19, Apr. 1998.

[19] M. J. Canty, *Image Analysis, Classification and Change Detection in Remote Sensing: With Algorithms for Python*, 4th ed. Boca Raton: CRC Press, 2023.

[20] N. Otsu, "A Threshold Selection Method from Gray-Level Histograms," *IEEE Transactions on Systems, Man, and Cybernetics*, vol. 9, no. 1, pp. 62–66, Jan. 1979.

[21] L. Bruzzone and D. Prieto, "Automatic analysis of the difference image for unsupervised change detection," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 38, no. 3, pp. 1171–1182, 2000.

[22] L. Kondmann, A. Toker, S. Saha, B. Scholkopf, L. Leal-Taixe, and X. X. Zhu, "Spatial Context Awareness for Unsupervised Change Detection in Optical Satellite Images," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 60, pp. 1–15, 2022.

[23] W. Dong, "What is OpenCV's INTER_AREA Actually Doing?" Jun. 2018, accessed : 19 June 2024. [Online]. Available: https://medium.com/@wenrudong/what-is-opencvs-inter-area-actually-doing-282a626a09b3

[24] L. Manhui, "Bobholamovic/ChangeDetectionToolbox," Feb. 2021, accessed : 22 June 2024. [Online]. Available: https://github.com/Bobholamovic/ChangeDetectionToolbox

[25] D. Berrar, "Performance Measures for Binary Classification," in *Encyclopedia of Bioinformatics and Computational Biology*. Elsevier, 2019, pp. 546–560.

[26] A. P. Cracknell, "Review article Synergy in remote sensing-what's in a pixel?" *International Journal of Remote Sensing*, Jan. 1998.

[27] J. C. Valdiviezo-N, A. Téllez-Quiñones, A. Salazar-Garibay, and A. A. López-Caloca, "Built-up index methods and their applications for urban extraction from Sentinel 2A satellite data: Discussion," *Journal of the Optical Society of America A*, vol. 35, no. 1, p. 35, Jan. 2018.

[28] R. Caye Daudt, B. Le Saux, A. Boulch, and Y. Gousseau, "OSCD - Onera Satellite Change Detection," Oct. 2019, under the CC BY 4.0 license, accessible at: https://creativecommons.org/licenses/by/4.0/.

[29] National Space Council Users' Advisory Group Climate and Societal Benefits Subcommittee, "Space Data Ethics: The Next Frontier in Responsible Leadership," Jan. 2023. [Online]. Available: https://www.nasa.gov/wp-content/uploads/2024/02/white-paper-space-data-ethics-2023-12-01-final-002.pdf

Appendix A : Details on Exclusion of NaN Values

This appendix lists the image pairs that are fully excluded from the evaluation as a result of having NaN values for some metrics and resolutions. It concerns 16 images of the LEVIR-CD dataset that do not contain any "change" pixels for some or all resolutions. In such a case, the recall and F-score cannot be calculated as it would require division by zero.

Table 3 lists the first resolution with no "change" pixels for each excluded image pair. Because of the downsampling process, images that initially contain few scattered "change" pixels can have no "change" pixels in lower resolutions, and are then fully excluded from the evaluation.

Image pair	**First resolution with no "change" pixels**
test_60	0.5 m/px
test_62	0.5 m/px
test_64	0.5 m/px
test_65	0.5 m/px
test_66	0.5 m/px
test_99	0.5 m/px
test_108	0.5 m/px
test_117	0.5 m/px
test_122	0.5 m/px
test_125	0.5 m/px
test_59	8 m/px
test_61	8 m/px
test_128	8 m/px
test_88	8 m/px
test_95	8 m/px
test_99	8 m/px

Table 3: For transparency, this table details the excluded image pairs and the resolutions at which they first have no "change" pixels.

The image pairs were fully excluded since otherwise different sets of image pairs would be considered between metrics and resolutions, complicating the comparison. The exclusion of the image pairs of course has an impact on the resulting values and trends in metrics. The time constraints of the project made it not possible to conduct an additional analysis of the excluded non-NaN values to estimate the impact.

Appendix B : Usage of ChatGPT

This appendix provides context on the usage of ChatGPT for the project. It covers the exact prompts used, as well as to which extent they were taken into account during the project. Overall, ChatGPT was used in a limited manner. It was not used to write or re-write text.

9.1 Content Related Questions

9.1.1 CVA Band Standardization

The most impacting usage of ChatGPT was for argumentation about the CVA band standardization in the paper. The standardization was already decided upon, and the experiment was already conducted before asking the questions. The initial reasons for using the standardization were critical reflection on the topic area, as well as it also being used in an ESA software [17].

The references provided by ChatGPT were either not accessible online, or only mentioned (radiometric) normalization, not standardization. Therefore, no scientific paper could be found that uses the same setup. The answer of ChatGPT resulted in the inclusion of the following arguments: (1) equal contribution of different spectral bands; and (2) limiting effects of varying light conditions. The answers were critically reflected upon, and the limitations of only using standardization are discussed in the paper as well.

The following questions were asked, in the same order:

> "Suppose Change Vector Analysis is used with multiple image bands. Should the image bands be normalized (standardised)? In any case, can you provide me references for your answer that I can use to follow-up on the information?"
>
> "Can you provide more specific resources that tackle z-score standardization for CVA?"
>
> "It seems that relative radiometric normalization is something that is more regularly used. Is z-score standardization still valid to use, and can you provide references?"
>
> "Is z-score standardization also useful when considering varying atmospheric conditions in CVA?"
>
> "I rather meant, varying lightness/darkness conditions for the overall image."

9.1.2 Confirmation of OTSU with IR-MAD

ChatGPT was asked concerning the usage of OTSU with IR-MAD. However, *this was after the experiment was already conducted and the choice was made for already decided reasons*. Therefore, the answer is not used as reference, and it does not influence the content of the paper. It was asked more as brief confirmation, though it is included here for transparency and completion. ChatGPT argued that OTSU is suitable, though of course, the answer requires critical reflection.

> "Do you know the IR-MAD algorithm for change detection?"
>
> "Suppose the IR-MAD algorithm is used for bitemporal binary change detection. After it is applied, a thresholding algorithm/simple theshold should be applied to get a binary "change"/"no change" map. Do you think that the Otsu thresholding technique is suited for this?"

9.1.3 Anti-996-License

The following question was asked to gain an understanding of the context for the Anti-996-License. However, the answer has not been taken as reference for the final text. Only the actual license has been taken as final reference.

> "Do you know the Anti-996-License?"

9.2 Code

ChatGPT was used to set the edge colors of specific plots, namely the plots of the mean metric values. A few lines of code were taken from the answers, that being a loop iterating through the edges to set the colors individually. The following prompts were used:

> "I am using matplotlib in Python. Why is there an edge that is grey (like my grid) which I cannot change? I am also trying to set the edge of the figure to black, with the grid being grey. I am using `ax.patch.set_edgecolor`, however it only creates lines underneath the outline in grey."
>
> "Why does `ax.patch.set_edgecolor` not work?"

Additionally, it was used to have different y axes left and right of the correlation matrices, resulting in including a few lines of code referencing the following prompt:

> "Can I have different labels on the left and right side for a heatmap in Seaborn and Matplotlib?"

9.3 Formatting Latex

For formatting a figure spanning two columns in Latex, ChatGPT was taken into account. It resulted in the usage of the `stfloats` package. The prompt was as follows:

> "I have a Latex document with two columns, and want an image to appear on the bottom of the page spread over two columns. \begin{figure*}[b!] does not work, and it appearson the last page of the full document. What can I do to fix it."