Standardization of Performance Metrics for Benchmarking InSAR Parameter Estimation

Francescopaolo Sica, Rafael Dill, Michael Schmitt

Department of Aerospace Engineering, University of the Bundeswehr Munich, Germany
(francescopaolo.sica, michael.schmitt)@unibw.de

Keywords: Synthetic Aperture Radar, SAR, Interferometry, InSAR, Evaluation Metrics, Benchmarking.

Abstract

Interferometric Synthetic Aperture Radar (InSAR) is a well-established remote sensing technique that enables a precise monitoring of the Earth’s surface. Accurate estimation of InSAR parameters, such as phase and coherence, is critical for deriving meaningful geophysical information. However, there are many challenges in evaluating the performance of different estimation methods. This work aims to standardize performance evaluation metrics to ensure consistent and reliable results across different applications. We address this challenge by examining factors such as the nature of the data (real or simulated), the processing stage of the evaluation, the signal properties under test, and the nature of the metrics (quantitative and qualitative). We emphasize the need for a comprehensive evaluation framework that integrates multiple metrics to assess different properties of the interferometric data. Examples of both simulated and real data are provided to demonstrate performance evaluation. In addition, we outline various metrics and evaluation frameworks to facilitate benchmarking in InSAR parameter estimation.

1. Introduction

Interferometric Synthetic Aperture Radar (InSAR) has emerged as a key technology in remote sensing, revolutionizing our ability to monitor the Earth’s surface with unprecedented accuracy and detail (Moreira et al., 2013). InSAR makes it possible to measure surface topography, ground displacements and their evolution over time at fine temporal and spatial scales. It facilitates the monitoring of natural hazards such as earthquakes, volcanic eruptions and landslides, providing invaluable insights into their dynamics and helping to assess associated risks (Ferretti et al., 2011, Fornaro et al., 2015). It also plays a critical role in the inversion of parameters of the imaged scene, e.g., agriculture and forestry (Solberg et al., 2017, Pulella et al., 2020), or land cover classification (Sica et al., 2019).

At the heart of any InSAR technique is the accurate estimation of interferometric phase and coherence, which are the key parameters for extracting meaningful information from SAR data. The interferometric phase contains valuable information about the topography or relative displacement between the radar sensor and the observed surface, while the coherence indicates the quality of the interferometric measurements. Accurate estimation of these parameters is essential to achieve accurate InSAR derived products. Various estimation methods have been developed over the years, ranging from classical algorithms (Lee et al., 2003) via recent nonlocal methods (Deledalle et al., 2015, Sica et al., 2018), to the newest and most advanced Deep Learning (DL) based techniques (Sica et al., 2020). These methods aim to optimize the trade-off between detail preservation and denoising power to meet the diverse needs of InSAR applications. The choice of the estimation techniques has a profound impact on the accuracy of phase and coherence estimation, which directly affects the reliability of the derived geophysical information. Therefore, standardization of performance evaluation metrics for InSAR parameter estimation is imperative to evaluate the effectiveness of different methods and to ensure consistent and reliable results across different applications.

In the field of SAR processing, previous review works on SAR despeckling have faced the problem of finding a fair framework for evaluating different despeckling algorithms, as done in (Singh et al., 2021). Similarly, in (Dellepiane and Angiati, 2013) the authors highlight the difference between the perceived quality of SAR despeckled images with respect to statistical metrics and propose an alternative to the usual performance metrics. Despite being a very important topic, there is not much review work in this direction. In addition, while SAR despeckling has been more studied, this type of analysis is still lacking for InSAR data.

In this paper, we address the importance of standardizing performance evaluation metrics for InSAR parameter estimation. Section 2 highlights the challenges and the rationale behind this work. In Section 3 we discuss the various aspects involved in the design of a testbed for the evaluation of estimation methods. In the final version of the manuscript, we will also provide example data, evaluation metrics, and a comparison of state-of-the-art methods, ranging from the full spectrum of classical methods to the latest DL approaches. Finally, in Section 6 we summarize the presented work and draw the perspectives planned for the final version of the manuscript.

2. Challenges

Whether analyzing a single pair of SAR images or constructing time series datasets, the reliability of InSAR measurements depends heavily on the accuracy of phase and coherence estimation. Algorithms developed to this aim focus on improving various aspects of the InSAR signal, among which the most important are: spatial and vertical resolution, preservation of fine details, and effective noise suppression. Achieving improvements in all of these aspects simultaneously, or even partially, is a significant challenge and requires thorough testing to ensure optimal performance. However, assessing the effectiveness of these algorithms is not always straightforward due to the multifaceted nature of the problem and limitations of existing metrics. Many metrics used for performance evaluation do not comprehensively consider all critical aspects simultaneously,

leading to potentially misleading conclusions when used singularly. For example, measuring only the regularity of the signal, such as residual calculations, tend to favor methods that oversmooth the signal. Consequently, it is essential to complement such metrics with others that specifically measure the degree of detail preservation. In addition, the quality of the signal itself adds another layer of complexity to the evaluation process. In certain applications, particularly those that involve multiple processing steps after phase estimation, a smoother signal may actually be advantageous, facilitating subsequent processes, as for example phase unwrapping. Therefore, it is imperative to consider the downstream effects of signal quality on the entire InSAR processing chain when evaluating estimation methods performance.

3. Performance evaluations criteria

Assessing the quality of an estimation method must consider several aspects, including the characteristics of the InSAR signal under test and all the possible methodologies for the evaluation of these characteristics. We identified the following important aspects for the evaluation of InSAR estimation methods:

1. Nature of the data
2. Nature of Metrics
3. Stage in the Processing Chain
4. Evaluated Signal Properties

3.1 Nature of the data

Performance evaluation has always been divided between testing on simulated or real data.

- Simulated data: Evaluation on simulated data provides a controlled environment for assessing algorithm performance. Simulated datasets allow systematic manipulation of signal characteristics and noise levels, facilitating a detailed understanding of algorithm behavior under idealized conditions. In (Sica et al., 2020), the authors show how to obtain a large and varied dataset for simulated InSAR data.
- Real data: Evaluation on real data is essential to validate algorithm performance in practical scenarios. Real-world datasets capture the complexity and variability of natural phenomena, providing insight into the robustness and reliability of algorithms under different environmental conditions. At the same time, the lack of a noise-free reference doesn’t allow the use of the most common performance metrics.

3.2 Nature of Metrics

The primary classification between metrics can be done depending on their nature.

- Qualitative metrics: Visual inspection of estimated signals provides qualitative insight into algorithm performance. Visualization of interferograms and coherence maps, as well as their comparison with external references of various types, can provide an intuitive understanding of signal characteristics and help identify important image details as well as anomalies or artifacts.
- Quantitative metrics: These provide objective measures of algorithm performance and are the most meaningful metrics. They allow for rigorous comparison and benchmarking. Metrics such as Root Mean Square Error (RMSE), Signal to Noise Ratio (SNR), and Correlation Coefficients provide numerical assessments of estimation precision and accuracy.

3.3 Stage in the Processing Chain

Another important aspect for this assessment is at what stage the quality of the estimated data should be evaluated: immediately after parameter estimation or sometime later in the processing chain.

- Immediately after estimation: The evaluation immediately after the estimation step is the classic method used in methodology papers. It focuses on evaluating the actual output of the estimation algorithm. This provides insight into the ability of the algorithm to accurately capture the underlying signal properties without influence from subsequent processing steps.
- At any point in the processing chain: Evaluation at intermediate or final stages of the processing chain considers the cumulative effects of estimation and additional processing steps on the InSAR signal. This holistic approach reflects real-world processing scenarios and provides a more comprehensive assessment of algorithm performance. On the downside, further processing also reduces resolution, which may penalize estimation methods that better preserve resolution and fine detail.

3.4 Evaluated Signal Properties

The metrics can be distinguished according to the characteristic of the signal that has to be tested. This is specifically true when using real data, which cannot rely on a noise-free reference.

- Signal regularity: these metrics evaluate the smoothness and continuity of the estimated signal. Algorithms that produce smoother outputs may be preferred in certain applications, while others may require preservation of finer details.
- Detail preservation: these metrics quantify the ability of the algorithm to preserve high-frequency components and fine-scale features in the estimated signal. This property is particularly important in applications where subtle variations or small-scale structures are of interest.

4. Sample data

In this section, we provide examples for synthetic and real data and highlight their suitability for a given performance evaluation.

4.1 Simulated data

Simulated data is probably the most commonly used data type for benchmarking InSAR parameter estimation methods. Historically, this has been due in part to the limited availability of real data, necessitating the simulation of certain situations and conditions. In addition, simulated data allow algorithm testing

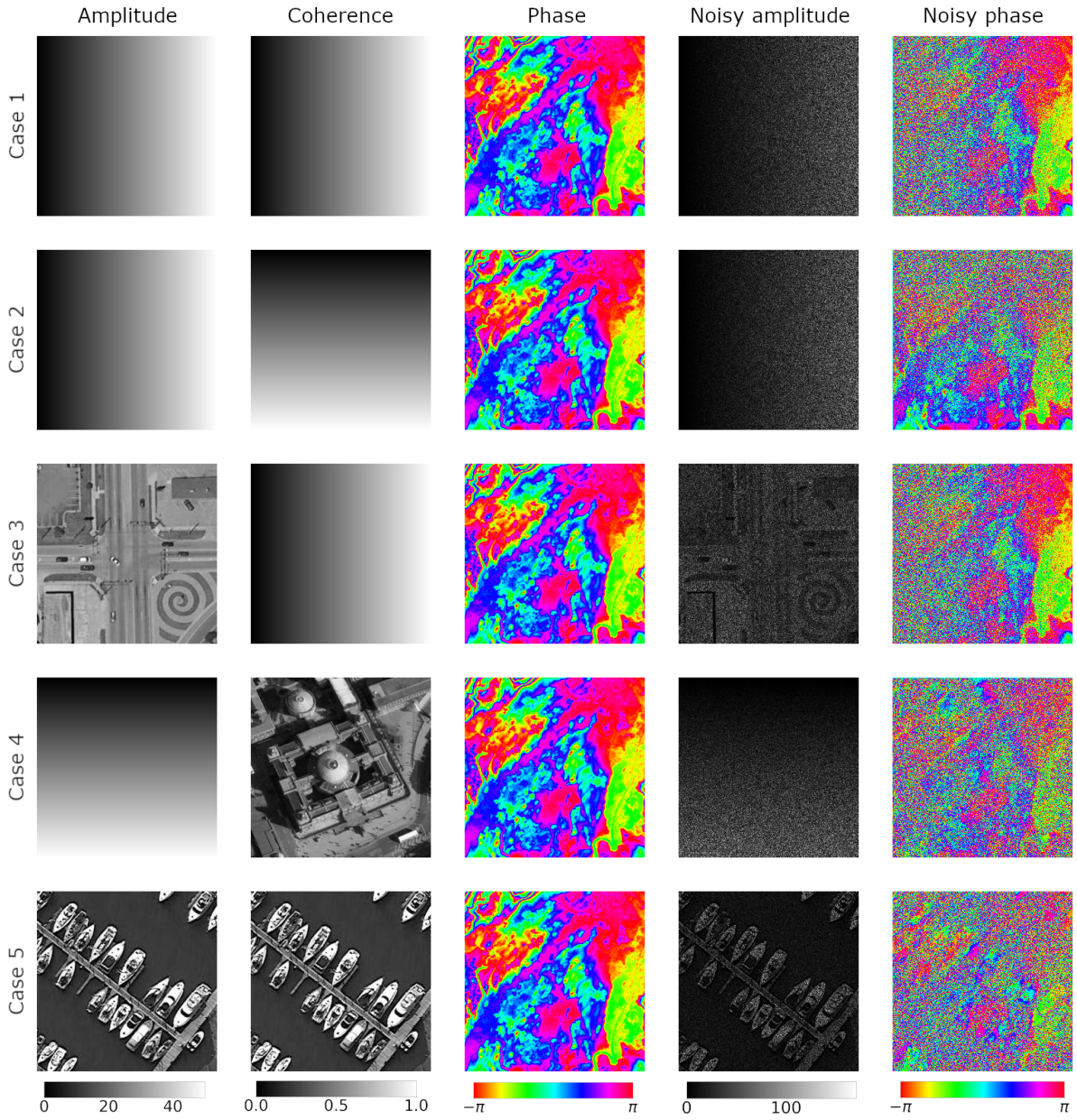

Figure 1. The five considered cases show all the possible combinations of noise-free parameters and the corresponding noisy data for low phase spatial variations.

in a controlled environment where the baseline (or reference) is perfectly known, commonly referred to as ground truth in remote sensing.

For simulated data to be effective, it must be generated reflecting real-world conditions as closely as possible. It should be designed to test specific properties of the algorithm under investigation that mirror those expected in real data. Therefore, the generation of simulated data must be done with critical attention to detail. In the context of training machine learning estimation algorithms, the authors in (Sica et al., 2020) introduced a methodology to simulate the main realistic scenarios that can be found in real data. The main aspects to consider in generating realistic simulated data are:

- Creation of a good image prior, i.e., the characteristics of the noise-free images for the three parameters.
- Generation of synthetic noise based on the InSAR statistical distribution (Sica et al., 2018).

Following this criteria, we provide in the following an example of simulated InSAR data priors generated by considering the variation of single parameters (amplitude, coherence, and phase) and their interdependence, together with the corresponding generated noisy data. Specifically, the amplitude and coherence may exhibit slow-varying trends or spatial patterns, while the phase may have smooth or abrupt variations, such as fringe patterns.

In Figure 1 and Figure 2, we show all possible combinations among noise-free parameters and the resulting noisy data for cases of slow and high phase spatial variation, respectively. In these figures, the first three columns show the noise-free

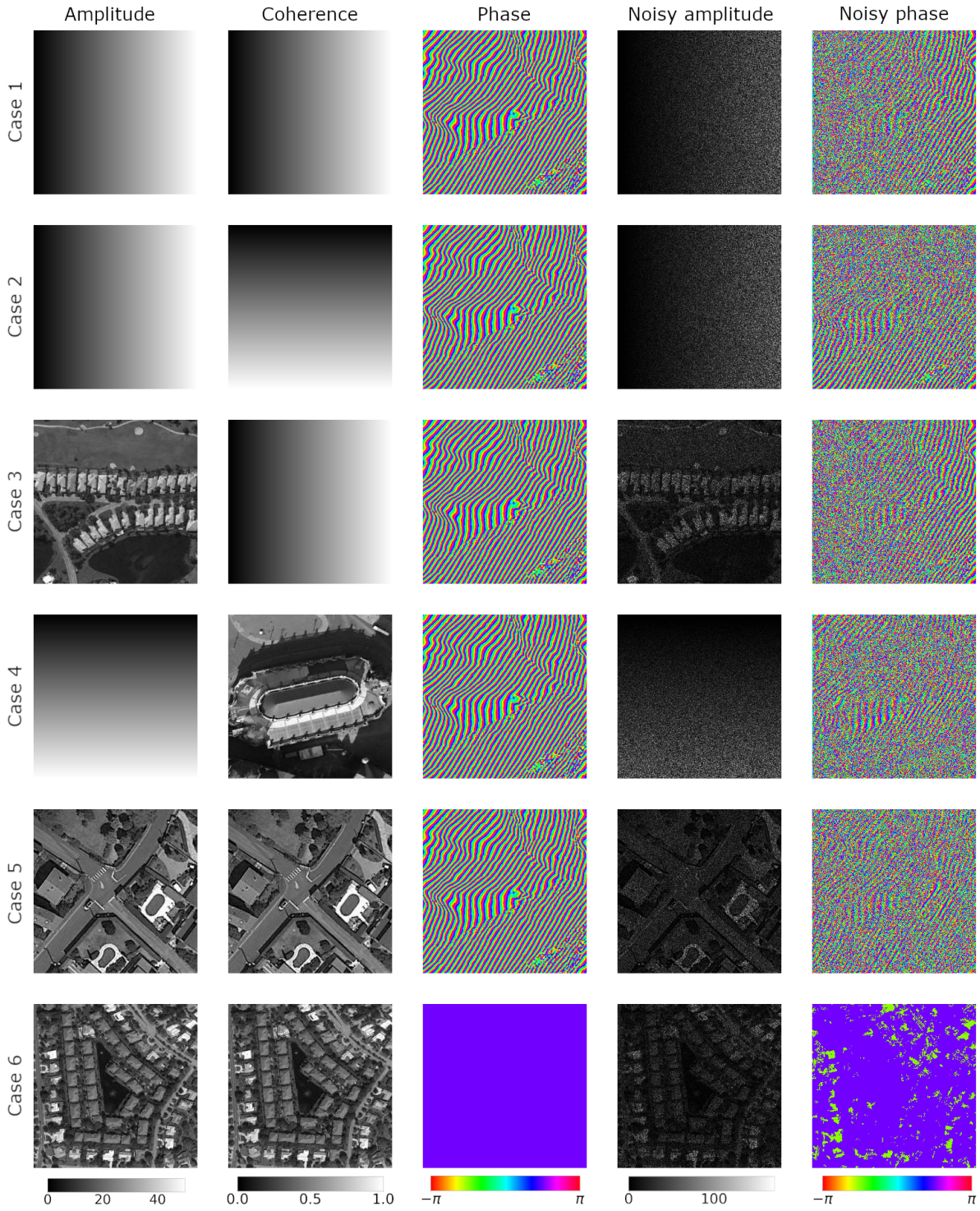

Figure 2. The six considered cases show all the possible combinations of noise-free parameters and the corresponding noisy data for high phase spatial variations.

amplitude, coherence, and phase, respectively. The last two columns show the corresponding noisy amplitude and noisy phase. Each row represents different combinations of these noise-free parameters. These controlled variations of the noise-free parameters allow to evaluate the ability of the algorithm to accurately estimate the InSAR parameters under different acquisition scenarios.

4.2 Real data

Today, an increasing amount of real data is available, making it valuable to evaluate algorithms on practical applications. However, the main limitation of using real data is the lack of a reference or ground truth, and therefore different evaluation strategies should be used.

Figure 3 shows an example of real data over a mountainous re-

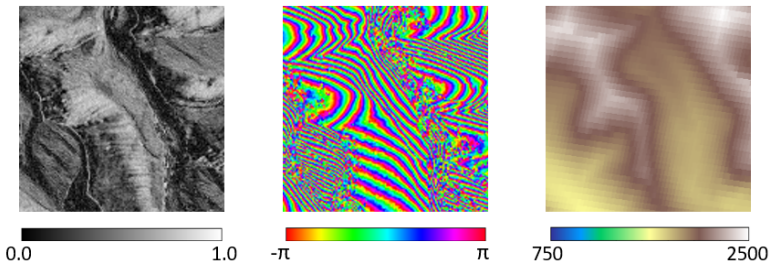

Figure 3. Example of real data of coherence, phase and DEM over a mountainous region.

gion. From left to right, the figure shows the coherence and phase estimated from the real data and the corresponding Digital Elevation Model (DEM). This challenging scenario contains areas of extremely dense fringe that were not included in the simulated data. Thus, it serves as an excellent test bed to provide additional challenging data for the evaluation of InSAR algorithms.

Possible benchmarking scenarios include comparing the estimated phase with the synthetic phase generated by an external DEM. Alternatively, the estimated phase can be unwrapped and correlated with the DEM. Finally, the unwrapped phase can be used to generate an elevation map that can be compared to an existing DEM.

5. Performance metrics

To evaluate the quality of the estimation algorithms, the subsequent performance metrics, as partially listed in (Vitale et al., 2022), were employed to ensure a reliable assessment across various applications. Starting with the mean squared error (MSE) and root mean squared error (RMSE) which are defined as

$$\text{MSE} = \frac{1}{m \times n} \sum_{i=1}^{m} \sum_{j=1}^{n} \left(\Phi(i,j) - \hat{\Phi}(i,j) \right)^2 \tag{1}$$

$$\text{RMSE} = \sqrt{\text{MSE}} \tag{2}$$

where, $\Phi(i,j)$ represents the value of the pixel at position (i,j) in the original interferogram, while $\hat{\Phi}(i,j)$ denotes the value of the corresponding pixel in the noisy interferogram. The variables m and n indicate the number of rows and columns (pixels) in the interferograms, respectively. The formula calculates the average of the squared differences between the corresponding pixel values in the original and noisy interferograms, thus quantifying the mean square error.

The structural similarity index (SSIM) is a metric used to measure the similarity between two images. As defined in (Wang et al., 2004), it evaluates changes in structural information, luminance, and contrast, providing a more perceptually relevant assessment compared to traditional metrics like MSE. Despite the fact that this metric is based mainly on a perceptual assessment, it can still provide information about the preservation of relevant structures in the data and can complement pixel-wise difference metrics, such as the MSE.

$$\text{SSIM}(x,y) = \frac{(2\mu_x\mu_y + c_1)(2\sigma_{xy} + c_2)}{(\mu_x^2 + \mu_y^2 + c_1)(\sigma_x^2 + \sigma_y^2 + c_2)} \tag{3}$$

where x and y represent the two images being compared, with

$$\mu_x = \frac{1}{N} \sum_{i=1}^{N} x_i \tag{4}$$

and analogously μ_y denoting their respective mean values, and N representing the number of pixels. The variables

$$\sigma_x = \sqrt{\frac{1}{N-1} \sum_{i=1}^{N} (x_i - \mu_x)^2} \tag{5}$$

and analogously σ_y being the standard deviation of the images x and y, while

$$\sigma_{xy} = \frac{1}{N-1} \sum_{i=1}^{N} (x_i - \mu_x)(y_i - \mu_y) \tag{6}$$

is the covariance between them. The constants c_1 and c_2 are used to stabilize the division with weak denominator values.

Pratt’s figure of merit (Pratt, 2001) abbreviated as FOM, is a measure used to evaluate the accuracy of edge detection algorithms by comparing detected edges to actual edges

$$\text{FOM} = \frac{1}{\max(N_A, N_D)} \sum_{i=1}^{N_D} \frac{1}{1 + \alpha \cdot d_i^2} \tag{7}$$

with N_A being the number of actual edge points, and N_D represents the number of detected edge points. The variable d_i denotes the Euclidean distance between the i-th detected edge point and the nearest actual edge point. The parameter α is a scaling factor that controls the sensitivity of the FOM to the distance d_i. The maximum function, $\max(N_A, N_D)$, ensures normalization by the greater number of actual or detected edge points. The summation accounts for all detected edge points, adjusting the contribution of each point based on its proximity to the nearest actual edge point.

The Kullback-Leibler divergence, as defined in (Kullback, 1978), computes the distance between the statistical distribution of the simulated noise and that of the filtered one

$$D_{KL}(P \parallel Q) = \sum_i P(i) \log \frac{P(i)}{Q(i)} \tag{8}$$

where Q represents the probability density function (pdf) of the predicted noise, and P represents the pdf of the simulated noise. An ideal filter will produce a $D_{KL} = 0$, indicating no divergence between the distributions of the simulated and filtered noise.

Finally, we also consider the *residues* evaluation metric, which doesn't require a ground truth reference and is therefore a good alternative to the performance metrics previously presented. Local phase jumps greater than π will lead to local errors that will affect the unwrapping procedure after phase estimation. Therefore, this measure provides information about the regularity of the signal and hence the quality of the estimation. These errors can be caused by steep terrain slopes or phase noise. These local errors can be detected by calculating the sum of the gradients in closed loops between four neighbouring pixels (Bamler and Hartl, 1998)

$$r(i,k) = \nabla \times \nabla\hat{\varphi}(i,k) = \nabla \times \mathbf{n}\nabla(i,k) \tag{9}$$

where:

- $r(i,k)$ represents the residue at the position (i,k).
- $\hat{\varphi}(i,k)$ is the estimated phase at position (i,k).
- $\nabla \times \nabla\hat{\varphi}(i,k)$ indicates the curl of the gradient of the estimated phase at (i,k).
- $\mathbf{n}\nabla(i,k)$ represents the solenoidal part of the phase gradient estimate field.
- $\nabla \times \mathbf{n}\nabla(i,k)$ denotes the curl of the solenoidal part of the phase gradient estimate field.

A solenoidal vector field, also known as a *divergence-free vector field*, is one where the divergence of the field is zero at every point. This means that the field lines are closed loops or extend infinitely without converging or diverging. Therefore the residue field can be expressed as

$$\begin{aligned} r(i,k) = \Delta\{\partial_i\varphi(i,k)\} + \Delta\{\partial_k\varphi(i+1,k)\} \\ - \Delta\{\partial_i\varphi(i,k+1)\} - \Delta\{\partial_k\varphi(i,k)\} \end{aligned} \tag{10}$$

where:

- $\varphi(i,k)$ is the phase at position (i,k),
- $\Delta\{\partial_i\varphi(i,k)\}$ is the difference operator (gradient) of the phase in the i-direction at position (i,k),
- $\Delta\{\partial_k\varphi(i,k)\}$ is the difference operator (gradient) of the phase in the k-direction at position (i,k),

Summing up all four difference operators of (10) results in the mentioned curl, which, if not zero, represents the detection of the residue at the given position.

6. Conclusions

In this paper, we have addressed several issues that arise in the evaluation of InSAR parameter estimation methods. We have considered factors such as the type of data (real or simulated), the processing stage, the signal properties being tested, and the type of metrics (quantitative and qualitative). By examining these aspects, we have emphasized the importance of a comprehensive evaluation that combines different metrics to simultaneously assess different properties of the interferometric data under test. We also provided examples of simulated and real data that can be used for performance evaluation, and described usable metrics and evaluation frameworks to facilitate benchmarking in the context of InSAR parameter estimation.

References

Bamler, R., Hartl, P., 1998. Synthetic aperture radar interferometry. *Inverse Problems*, 14(4), R1.

Deledalle, C.-A., Denis, L., Tupin, F., Reigber, A., Jäger, M., 2015. NL-SAR: A unified nonlocal framework for resolution-preserving (Pol)(In) SAR denoising. *IEEE Transactions on Geoscience and Remote Sensing*, 53(4), 2021–2038.

Dellepiane, S. G., Angiati, E., 2013. Quality assessment of despeckled SAR images. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 7(2), 691-707.

Ferretti, A., Fumagalli, A., Novali, F., Prati, C., Rocca, F., Rucci, A., 2011. A new algorithm for processing interferometric datastacks: SqueeSAR. *IEEE Transactions on Geoscience and Remote Sensing*, 49(9), 3460–3470.

Fornaro, G., Verde, S., Reale, D., Pauciullo, A., 2015. CAESAR: An approach based on covariance matrix decomposition to improve multibaseline-multitemporal interferometric SAR processing. *IEEE Transactions on Geoscience and Remote Sensing*, 53(4), 2050–2065.

Kullback, S., 1978. *Information Theory and Statistics*. Dover Publications, Inc., Gloucester, Massachusetts.

Lee, J. S., Cloude, S. R., Papathanassiou, K. P., Grunes, M. R., Woodhouse, I. H., 2003. Speckle filtering and coherence estimation of polarimetric SAR interferometry data for forest applications. *IEEE Transactions on Geoscience and Remote Sensing*, 41(10), 2254-2263.

Moreira, A., Prats-Iraola, P., Younis, M., Krieger, G., Hajnsek, I., Papathanassiou, K. P., 2013. A tutorial on synthetic aperture radar. *IEEE Geoscience and remote sensing magazine*, 1(1), 6-43.

Pratt, W. K., 2001. *Edge Detection*. John Wiley Sons, Inc., New York, 490–494.

Pulella, A., Aragão Santos, R., Sica, F., Posovszky, P., Rizzoli, P., 2020. Multi-temporal sentinel-1 backscatter and coherence for rainforest mapping. *Remote Sensing*, 12(5), 847.

Sica, F., Cozzolino, D., Zhu, X. X., Verdoliva, L., Poggi, G., 2018. InSAR-BM3D: A nonlocal filter for SAR interferometric phase restoration. *IEEE Transactions on Geoscience and Remote Sensing*, 56, 3456-3467.

Sica, F., Gobbi, G., Rizzoli, P., Bruzzone, L., 2020. -Net: Deep residual learning for InSAR parameters estimation. *IEEE Transactions on Geoscience and Remote Sensing*, 59, 3917-3941.

Sica, F., Pulella, A., Nannini, M., Pinheiro, M., Rizzoli, P., 2019. Repeat-pass SAR interferometry for land cover classification: A methodology using Sentinel-1 Short-Time-Series. *Remote Sensing of Environment*, 232, 111-277.

Singh, P., Diwakar, M., Shankar, A., Shree, R., Kumar, M., 2021. A Review on SAR Image and its Despeckling. *Archives of Computational Methods in Engineering*, 28, 4633-4653.

Solberg, S., Hansen, E. H., Gobakken, T., Naessset, E., Zahabu, E., 2017. Biomass and InSAR height relationship in a dense tropical forest. *Remote Sensing of Environment*, 192, 166-175.

Vitale, S., Ferraioli, G., Pascazio, V., Schirinzi, G., 2022. InSAR-MONet: Interferometric SAR Phase Denoising Using a Multiobjective Neural Network. *IEEE Transactions on Geoscience and Remote Sensing*, 60, 1-14.

Wang, Z., Bovik, A., Sheikh, H., Simoncelli, E., 2004. Image quality assessment: from error visibility to structural similarity. *IEEE Transactions on Image Processing*, 13(4), 600-612.