RESEARCH ARTICLE / ARTÍCULO DE INVESTIGACIÓN
https://dx.doi.org/10.14482/inde.44.01.985.544

Threshold-Based Identification of non-Gaussian Distortion in Optical Constellations Using Clustering Validity Metrics

Identificación basada en umbrales de la distorsión no gaussiana en constelaciones ópticas usando métricas de validez de agrupamiento

Eduardo Avendaño Fernández*
Henry Montaña Quintero**
Lely A. Luengas-C.***

* Titular Professor, Universidad Pedagógica y Tecnológica de Colombia (UPTC), Facultad Seccional Sogamoso, Sogamoso (Colombia). PhD. Orcid-ID: https://orcid.org/0000-0003-0910-8539. eduardo.avendano@uptc.edu.co

** Associate Professor, Universidad Distrital Francisco José de Caldas (UDFJC), Facultad Tecnológica, Bogotá, DC. (Colombia). Magister. Orcid-ID: https://orcid.org/0000-0003-0752-6315. hmontanaq@udistrital.edu.co

*** Titular Professor, Universidad Distrital Francisco José de Caldas (UDFJC), Facultad Tecnológica, Bogotá, DC. (Colombia). PhD. Orcid-ID: https://orcid.org/0000-0002-3600-4666. laluengasc@udistrital.edu.co

Correspondence: Eduardo Avendaño Fernández, Sogamoso, Boyacá (Colombia). Office Phone: +57 60 7706740 Ext: 2621.

Abstract

This work presents a threshold-based demodulation strategy for 16-QAM and 4+12 PSK constellations impaired by non-Gaussian distortions. The proposed method uses clustering validity indices as a decision metric. By applying fragmentation through clustering algorithms –k-means, fuzzy c-means (FCM), and Gustafson-Kessel FCM (GK-FCM)– we were able to identify ellipsoidal distortions on external data-symbol clusters and dynamically select appropriate demodulation strategies. The proposed clustering-based approach does not require IQ (in-phase and quadrature) branch imbalance and phase-offset corrections by redefining decision regions based on cluster centroids. We introduce the use of clustering validity indexes (Partition Coefficient, Separation, Xie and Beni's, and Dunn Index) to characterize symbol distortion levels in constellation diagrams and establish performance thresholds. The combination of DI and XB provides a criterion for defining the threshold of non-Gaussian distortion. In particular, XB ≥ 10.7 and DI ≥ 0.015 may serve as empirical indicators that the constellation in radio over fiber (RoF) optical systems has transitioned into a more structured regime where the cluster centroids are used for demodulation. Experimental results show that at high-noise levels (the optical signal to noise ratio OSNR = 16 dB), the XB index reaches its minimum value, confirming the method's sensitivity to noise-induced distortion. Improvements in the optical signal-to-noise ratio (OSNR) of up to 2.1 dB for 16-QAM and 0.7 dB for 4+12 PSK were observed at a BER threshold of 10^{-2} after transmission over 78.8 km of fiber. The combination of DI and XB indices provides a robust criterion for defining the threshold of non-Gaussian distortion. These experimental findings suggest that clustering validity metrics can serve as effective thresholds for adaptive demodulation, enabling real-time identification of non-Gaussian distortions in RoF communication systems.

Keywords: clustering, k-means, fuzzy c-means, Gustafson-Kessel, non-Gaussian distortion, nonlinear phase noise.

Fecha de recepción: 17 de julio de 2025
Fecha de aceptación: 4 de agosto de 2025

Resumen

En este trabajo, los autores demuestran experimentalmente una estrategia de demodulación basada en umbrales para constelaciones 16-QAM y 4+12 PSK afectadas por distorsiones no gaussianas, utilizando índices de validez de agrupamiento como métrica de decisión. Al aplicar fragmentación mediante algoritmos de clustering –k-means, fuzzy c-means (FCM) y Gustafson-Kessel FCM (GK-FCM)– lograron identificar distorsiones elipsoidales en los clústeres externos de símbolos de datos y seleccionar dinámicamente estrategias de demodulación apropiadas. El enfoque propuesto, basado en agrupamiento, no requiere correcciones de desbalance en las ramas IQ (en fase y cuadratura) ni de desfase, ya que redefine las regiones de decisión en función de los centroides de los clústeres. Los autores introducen el uso de índices de validez de agrupamiento (Coeficiente de Partición, Separación, Xie y Beni, y el Índice de Dunn) para caracterizar los niveles de distorsión de los símbolos en los diagramas de constelación y establecer umbrales de desempeño. La combinación de DI y XB proporciona un criterio para definir el umbral de distorsión no gaussiana. En particular, XB ≥ 10.7 y DI ≥ 0.015 pueden servir como indicadores empíricos de que la constelación en sistemas ópticos de radio sobre fibra (RoF) ha transitado hacia un régimen más estructurado en el que los centroides de clúster son utilizados para la demodulación. Los resultados experimentales muestran que, en condiciones de alto ruido (la relación señal a ruido óptica) OSNR = 16 dB, el índice XB alcanza su valor mínimo, confirmando la sensibilidad del método a la distorsión inducida por ruido. Se observaron mejoras en la relación señal-ruido óptica (OSNR) de hasta 2.1 dB para 16-QAM y 0.7 dB para 4+12 PSK en un umbral de BER de 10^{-2} en una transmisión sobre 78.8 km de fibra. La combinación de los índices DI y XB proporciona un criterio sólido para definir el umbral de distorsión no gaussiana. Estos hallazgos experimentales sugieren que las métricas de validez de agrupamiento pueden servir como umbrales efectivos para la demodulación adaptativa, permitiendo la identificación en tiempo real de distorsiones no gaussianas en sistemas de comunicación RoF.

Palabras clave: agrupamiento, k-means, fuzzy c-means, Gustafson-Kessel, distorsión no-gaussiana, ruido de fase no lineal.

INTRODUCTION

Radio-over-Fiber (RoF) has emerged as a key enabler for next-generation wireless networks, offering a cost-effective and high-capacity solution for the seamless integration of optical and wireless domains. By transporting radiofrequency (RF) signals over optical fiber, RoF supports centralized processing architectures such as Cloud-RAN (Radio Access Networks), which are essential for ultra-dense small cell deployments in 5G and beyond. The low-loss and wide bandwidth characteristics of optical fiber make RoF particularly suitable for high-frequency millimeter-wave and sub-THz bands, where wireless propagation suffers from severe attenuation. Furthermore, RoF simplifies the distribution of massive multiple input multiple outuput (MIMO) and beamforming signals, reducing the complexity and cost of remote antenna units. Its compatibility with existing fiber infrastructure facilitates rapid deployment while providing a unified fronthaul, midhaul, and backhaul platform to meet stringent latency and reliability requirements for applications like ultra-reliable low-latency communications (URLLC) [1] and industrial Internet of Things (IoT). By enabling multi-band and multi-standard transmission over a single optical link, RoF supports the convergence of heterogeneous networks, ensuring scalable and future-proof connectivity for 6G and beyond. The next-generation fiber-wireless communications systems require flexible receiver architectures able to process any data rate associated with changing m-ary modulation formats according to channel state information. The simplest technique to generate and distribute radiofrequency (RF) modulated signals over fiber, commonly referred to as Radio over Fiber (RoF) architectures, is the conventional Intensity Modulation/Direct Detection (IM/DD) [2]. However, it suffers from distortions due to the intrinsic nonlinear characteristics of the external Mach-Zehnder Modulator (MZM) needed in the electrical-to-optical conversion, the interaction of optical fiber dispersive effects [3], along with frequency mismatch between the RF received signal and the local oscillator (LO), after the optical detection stage. Moreover, several sources of noise (laser noise, shot noise, thermal noise, among others) and other kinds of impairments, such as harmonic distortion and intermodulation distortion, limit the dynamic range and decrease the system performance of the RoF system [4]. Detailing a specific type of distortion, only the fluctuations of the power and the optical phase, plus the use of single-frequency lasers that do not exhibit a perfect sinusoidal oscillation of the electric field at their output, give rise to a nonlinear phenomenon known as phase noise (PN) [5]. Besides, the joint effects of laser phase fluctuation and the fiber dispersion introduce a mismatch between the optical carrier and the microwave/millimeter-wave signals. The PN in the RoF studied scenario is introduced by the mismatch between the optical carrier and the transmitted radiofrequency signals that are not correlated within the coherence time [6]. Therefore, this paper focuses on the impact of residual phase noise (PN) on symbol constellations in intensity modulation/direct detection (IM-DD) radio-over-fiber (RoF) systems, where phase recovery is not performed but residual PN and dispersion-induced effects in the optical and electrical domains still impair symbol quality, and few mitigation strategies have been reported in the state of the art. However, the demodulation methods take into account all distortions introduced by system devices at the transmitter and receiver, and the channel in the end-to-end path. The PN impact on the symbols degrades the error vector magnitude (EVM) and causes phase error, as a function of fiber transmission distance. The impact of dispersion induced by PN on signals transmitted over Single Sideband (SSB) modulation RoF systems was characterized experimentally in [7] we propose an analog multiple intermediate-frequency-over-fiber (multi-IFoF, [8], and a phase adjustment technique based on an optical spectrum processor enables the suppression of this effect. Several feed-forward algorithms have been validated as effective methods to mitigate the phase fluctuation of the laser sources. In a comparative study on three Carrier Phase Estimation (CPE) algorithms [9], including one-tap normalized least mean square (NLMS), the block average method, and the Viterbi-Viterbi algorithm, the NLMS exhibits an acceptable performance at the price of hard step-size optimization. On the other hand, the authors in [10] propose a two-stage extended Kalman filtering (EKF) technique for the joint compensation of frequency offset (FO), linear and nonlinear phase noise, and amplitude noise in QAM systems. In the first stage, a coarse compensation of FO is performed using a set of training data symbols, and in the second stage, using the EKF, a fine compensation of the residual PO, PN due to laser linewidth and nonlinear effects is performed at the cost of higher computational effort.

Considering small improvements on threshold limits and that equalization and optimization algorithms impose a strong trade-off in computational cost, some machine learning approaches have been applied in optical communications. In [11] and [12], a support vector machine (SVM) classifier is introduced to create a nonlinear decision boundary in m-ary PSK-based coherent optical systems to mitigate nonlinear PN. The training is performed using binary labels for different SVMs according to the number of symbols in the constellation and reaches a maximum transmission distance of 480km using a launching power of 2 dBm, at a BER of 10-3. Moreover, the nonlinear decision boundary can be flexibly adjusted to create an irregular shape and enable more precise classification. Another machine learning detector based on the k-nearest neighbors (kNN) algorithm is proposed in [13] to overcome mainly PN and nonlinear PN in zero-dispersion links and dispersion-managed links. An improved algorithm referred to as distance-weight DW-kNN is introduced, and it outperforms the maximum likelihood (ML) post-compensation approach. Additionally, due to the temporal variation introduced by mechanical perturbations in optical fiber, temperature oscillation, and bias drift of the IQ components in the modulators, a non-Gaussian distribution in data-symbol points may shift the centroids of the received symbols from their ideal constellation positions. To mitigate these impairments, an adaptive machine learning-based non-symmetrical decision technique proposed in [14] studies the time-varying impairments in a 16-QAM Nyquist system at 16 GBd in back-to-back and 250 km links.

As reviewed before, there is a small gap for improving performance where flexible techniques enabled by machine learning and artificial intelligence algorithms may take the lead. Furthermore, considering the joint effects of system devices and the optical fiber phenomena during propagation, our contribution to the state of the art resides in the introduction of constellation fragmentation by clustering and robust signal demodulation (identifying non-Gaussian distortion), including noise characterization over any modulated signal to minimize bit errors after de-mapping [15]. For testing the method, we have implemented a RoF system due to being a well-established and cost-effective technology that exhibits channel transmission impairments of both wireless and optical fiber domains. The results for two modulation schemes, 16-QAM and 4+12 PSK, show better performance compared to the conventional demodulation technique.

Recent studies have proposed adaptive demodulation architectures based on clustering and deep learning to mitigate nonlinear and non-Gaussian impairments more effectively [16]-[17]. Moreover, [18] highlights the importance of real-time metrics to distinguish non-Gaussian behavior in optical channels. Our contribution addresses this challenge by defining threshold-based metrics derived from clustering validity indices to dynamically identify and compensate non-Gaussian distortions.

The remainder of this paper is organized as follows: in Section 2, an introduction to clustering algorithms and validation indices is explained for signal demodulation; then, in Section 3, the Radio-over-Fiber experimental setup is presented; in Section 4, the analysis of results and discussions are shown; and finally, conclusions and future work are summarized in Section 5.

METHODOLOGY

Clustering Techniques for non-Gaussian Distortion Identification and Demodulation

Clustering

It is a type of unsupervised machine learning technique used to find homogeneous subgroups from a data set X, such that objects in the same group (clusters) are more similar than those in other groups. In the context of optical and digital communications, the clustering methods are used to properly classify each point of a data-symbol constellation, which is scattered due to various imperfections of system devices and physical phenomena of optical fiber arising during transmission. Three clustering algorithms are explained as follows:

k-means

It is a widely used clustering algorithm applied for phase recovery and symbol demodulation in optical and RoF systems [19], [20]. For modulation formats such as 16-QAM and 4+12 PSK, it partitions the constellation into k clusters (e.g., 16 for 16-QAM), assigning each symbol to the nearest centroid based on Euclidean distance and iteratively updating centroid positions until convergence. Its main advantages are simplicity and fast convergence; however, it assumes spherical clusters, making it less effective under non-Gaussian distortions or ellipsoidal symbol deformation caused by phase noise or dispersion.

Fuzzy c-means (FCM)

It is a soft-clustering algorithm that assigns each symbol a degree of membership across multiple clusters, improving robustness in noisy or overlapping constellations [21]. It minimizes a weighted objective function where memberships and cluster centroids are iteratively updated, allowing finer separation of distorted symbols compared to hard clustering. For modulation formats such as 16-QAM, this flexibility helps handle moderate non-Gaussian effects, especially in outer symbols affected by phase noise or dispersion. However, its higher computational cost and reliance on isotropic distance metrics limit its effectiveness under severe ellipsoidal distortions, where covariance-adaptive approaches like GK-FCM become more suitable.

Gustafson-Kessel Clustering

The Gustafson-Kessel algorithm [22] associates each sample-point with its centroid and its covariance. The main feature of this clustering algorithm is its distance, which adapts to the shape of the cluster by estimating the cluster covariance matrix and adjusting the distance accordingly [23]. The objective function J_m of the FCM-GK algorithm is defined as

$$J_m = \sum_{i=1}^{c} \sum_{k=1}^{N} \mu_{ik}^{m} D_{ik}^{2} A_i \tag{1}$$

However, since this function is linear concerning the distance norm matrix Ai, direct minimization would lead to trivial, non-informative solutions (e.g., $A_i \to 0$). To ensure a meaningful optimization, A_i is constrained by fixing its determinant $|A_i|=1$, which preserves cluster volume while allowing the matrix to adapt its shape (covariance) to better capture ellipsoidal shape distortions in the symbol-data. This makes FCM–GK [24] particularly well-suited for the non-Gaussian, phase-noise-distorted clusters observed in RoF constellations [25]. Applied to demodulation in optical RoF systems, FCM–GK has proven to outperform k-means and soft FCM in handling anisotropic symbol deformation under transmission impairments, achieving OSNR gains up to ~2.9 dB for 16QAM and ~1.4 dB for 4+12 PSK without requiring separate IQ imbalance or phase-offset compensation.

Validation Indexes

A common approach for quantitatively evaluating a data partition is to use relative validity indexes, where each candidate partition obtained by a clustering algorithm is compared to other partitions of the same data set, making it possible to estimate the number of clusters from data. Therefore, for validation purposes, we assess different functions that provide cluster validity measures and goodness for the obtained partitions [24]. Different scalar validity measures have been proposed in the literature, but we focus on the indices described in the Fuzzy Clustering and Data Analysis Toolbox [26] as follows:

Partition Coefficient (PC): measures the amount of overlap between clusters. It is defined as

$$PC(c) = \frac{1}{N} \sum_{i=1}^{c} \sum_{j=1}^{N} \left(\mu_{ij}\right)^2 \tag{2}$$

Where μ_{ij} is the membership of the data-point j in the cluster i. The optimal number of clusters is the minimum value.

Partition Index (SC): is the ratio of the sum of compactness and separation of the clusters. It corresponds to the sum of individual cluster validity measures normalized by the fuzzy cardinality of each cluster.

$$SC(c)=\sum_{i=1}^{c}\frac{\sum_{j=1}^{N}\left(\mu_{ij}\right)^{m}\left\|x_{j}-v_{i}\right\|^{2}}{N_{i}\sum_{k=1}^{c}\left\|v_{k}-v_{i}\right\|^{2}} \tag{3}$$

SC index is useful when comparing different partitions with an equal number of clusters, and a lower value indicates a better partition.

Separation index (S): On the contrary of partition index (SC), the separation index uses a minimum-distance separation for partition validity. It is defined by

$$S(c)=\frac{\sum_{i=1}^{c}\sum_{j=1}^{N}\left(\mu_{ij}\right)^{2}\left\|x_{j}-v_{i}\right\|^{2}}{N\,min_{i,k}(\|v_{k}-v_{i}\|^{2})} \tag{4}$$

The Xie and Beni’s index (XB): represents a fuzzy-validity criterion based on a function that identifies overall compact and separate fuzzy c-partitions.

$$XB(c)=\frac{\sum_{i=1}^{c}\sum_{j=1}^{N}\left(\mu_{ij}\right)^{m}\left\|x_{j}-v_{i}\right\|^{2}}{N\min_{i,j}\left(\left\|x_{j}-v_{i}\right\|^{2}\right)} \tag{5}$$

The more separate the clusters, the larger the minimum distance between cluster centroids, minimizing the value of the XB index.

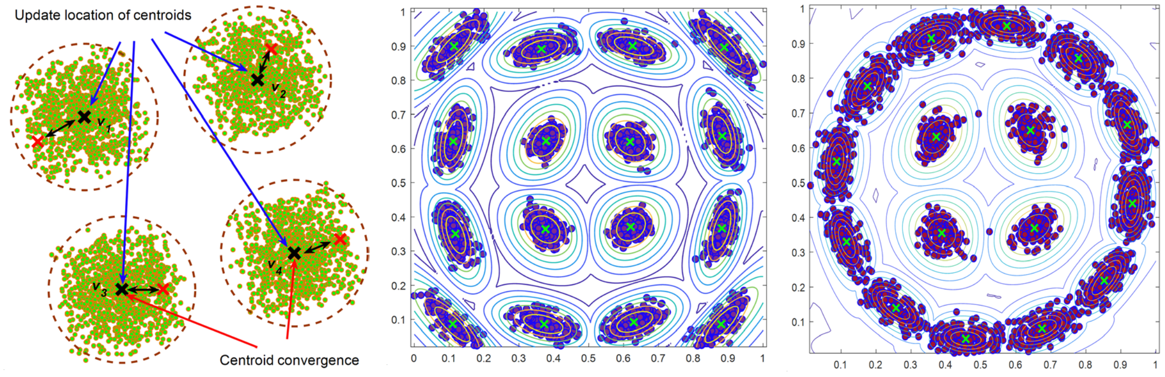
The Dunn index (DI): measures compactness (maximum distance in between data points of clusters) and cluster separation (minimum distance between clusters). The following equation indicates how to obtain the index:

$$DI(c) = \min_{i \in c} \left\{ \min_{j \in c, i \neq j} \left\{ \frac{\min_{x \in C_i, y \in C_j} d(x,y)}{\max_{x,y \in C} d(x,y)} \right\} \right\} \qquad (6)$$

All these cluster validation indices will be applied to determine the quality of a given clustering, and the computational effort will decrease, knowing the correct number of clusters, for the case of 16 symbols.

Demodulation of Non-Gaussian Constellations in Optical Communications

The general setting for the clustering-based demodulation (only FCM-GK is explained for simplicity; the procedure is similar for k-means and FCM) is applied as follows: i) the FCM-GK algorithm requires as inputs, the data-symbol observations and the number of clusters vi (predefined value of 16 for 16-QAM modulation format), and their expected centroids ideal locations of the 16-QAM or 4+12 PSK constellations. The data symbols correspond to bidimensional (2-D) vectors $x(k) = [x_i(k)\ x_q(k)]$, where x_i and x_q are the in-phase and quadrature (IQ) component samples projected on the complex plane. From this data-set, the weighted exponent $m = 2$ with a tolerance criteria $\varepsilon < 0.001$ as proposed in [27], and the partition matrix U are all initialized to avoid biasing the clustering process towards pre-defined centroids and to ensure convergence independent of prior knowledge of symbol positions; ii) the centroid of clusters with distortion are estimated as indicated in equation (3) and, as shown in Fig. 1a; iii) after that, membership values are obtained for each data concerning the closest centroid, calculating the distance norm (k-means and FCM uses Euclidean norm, and, FCM-GK uses Mahalanobis distance norm, but constrained to fixed determinant of A equal to 1); iv) the matrix fuzzy partition U is updated, and finally, v) the criterion for termination is calculated as $\|U(l)\text{-}U(l\text{-}1)1\| < \varepsilon$, if convergence is achieved, the algorithm stops; if not, it returns to step iii (as shown in Figure 1a with the center of mass being a marker ×). Figures 1b and 1c show the 16-QAM and 4+12 PSK constellations after the clustering process, including the update of centroids until convergence [25]. Besides, as observed in Figures 1b and 1c, the external symbols suffer non-Gaussian distortion while the inner symbols exhibit a quasi-circular shape.



1a) 1b) 1c)

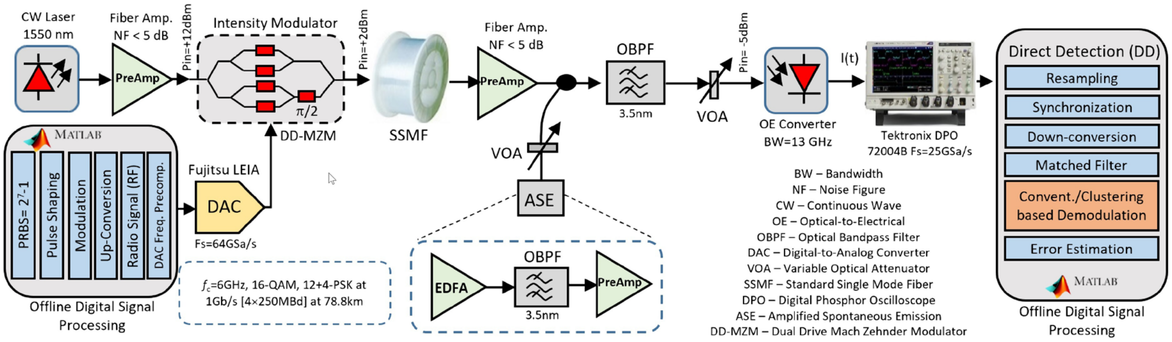
Source: own elaboration.

Figure 1. Clustering of data symbols: a) General clustering principle, b) clustering of 16-QAM, and c) clustering of 12+4 PSK

These effects degrade the performance when a conventional demodulation grid is used. Besides, the lower Euclidean distance among clusters for the 4+12 PSK constellation; for that reason, a higher BER will be introduced.

Experimental Setup

Figure 2 shows the schematic diagram for the RoF system setup as a single-channel system. A pseudorandom binary sequence (PRBS) with a length of 218 at a baud rate of 250MBaud was generated to deliver 1Gb/s in a single-channel and single-polarization case. The pulse shaping uses a Root Raised Cosine (RRC) filter with eight taps and a roll-off factor commonly used of α=0.2. After that, the modulation process was performed by encoding 4 bits per symbol (2^4 = 16 symbols) for both 16-QAM and 12+4 PSK. Then, the up-conversion stage translated the baseband signal to a radio frequency (RF) carrier at 6GHz. This signal was digitized through a Fujitsu LEIA board with a sampling frequency of 64 GSa/s and entered into the Mach Zehnder modulator RF inputs, and the RF signal was propagated over a spool of fiber with a length of 78.8 km. As an optical source, we used a distributed feedback (DFB) laser (linewidth 100kHz) emitting at 1550nm [28]. Amplified Spontaneous Emission (ASE) noise was injected and filtered by an optical bandpass filter (OBPF) with a bandwidth of 3.5 nm. The ASE noise is used to determine the functionality of the three clustering algorithms in the face of different OSNR conditions. A variable optical attenuator (VOA) was adjusted to set the power of -5 dBm at the input of the optical-to-electrical converter with a bandwidth up to 13GHz. Then, the electrical signal was entered into a digital oscilloscope at a sampling rate of 25 GSa/s, and the captured data was stored for offline processing. The down-conversion stage translates the RF signal to baseband, and then the matched filtering with the same prototype characteristics as the transmitter side filter is applied.



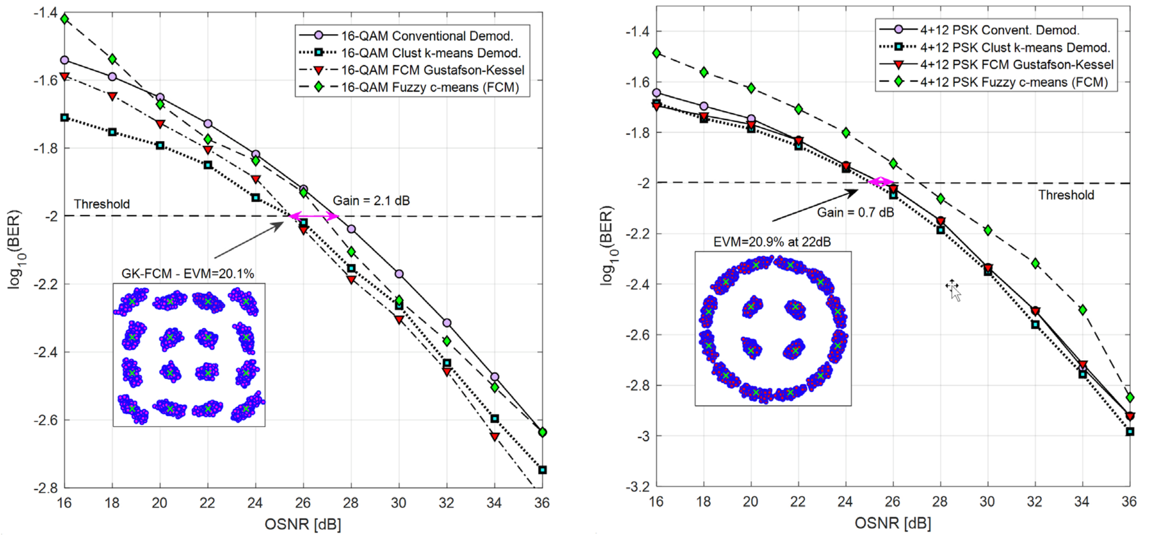
Source: own elaboration.

Figure 2. RoF experimental Setup

The demodulation stage, applying conventional rectangular grid demodulation, k-means, FCM, and GK-FCM clustering-based methods, is evaluated, varying the optical signal-to-noise ratio (OSNR) from 16 dB to 36 dB as a function of bit error rate (BER).

RESULTS AND DISCUSSION

The demodulated constellations exhibit strong non-Gaussian effects on clusters, due to the moderate adopted baud-rate, which translates to a long symbol duration. This symbol duration must be compared to the coherence time (t_c) between local oscillator and the incoming signal from the transmitter (ideally, a baud rate of 250 MBd needs a (t_c) around 4 ns). The clustering-based demodulation techniques mitigate the phase noise introduced by the mismatch between the optical carrier and the transmitted radiofrequency signals that are not correlated within the coherence time [11] of the local oscillator for the specific baud rate (250 MBd) used in the experimental setup.



3a)

3b)

Source: own elaboration.

Figure 3. BER performance at 1 Gb/s after 78.8 km span for a) 16-QAM and b) 4+12 PSK

Observing Figure 3a, the 16-QAM constellation shows higher separation (based on Euclidean distance metric) among the data-symbol points and a non-Gaussian effect, mainly in the corners of the external grid, where clusters with ellipsoidal shapes will overlap if the conventional grid is used. In Figure 3a, we plot the BER performance versus OSNR for 16-QAM. The continuous line with the circle marker corresponds to the conventional demodulation method. The error performance for FCM is shown with the dot-dashed line and a diamond marker; with a square marker, the dotted line shows the k-means algorithm. Finally, with a triangle marker and dot-dashed line, the GK-FCM method, the constellation in the inset shows the clusters and recovered symbols of a 16-QAM constellation with the FCM-GK method. k-means and GK-FCM perform a similar gain of 2.1 dB in the OSNR scale compared with conventional demodulation of 16QAM. However, beyond 26 dB, the gain margin for GK-FCM over k-means holds around 1dB for the rest of the curve up to 36 dB. The constellation inset shows the data-symbol distribution for GK-FCM, obtaining an error vector magnitude (EVM) of 20.1% for an OSNR of approximately 25 dB. However, the k-means algorithm performs better under low OSNR levels.

The proposed method is designed to identify constellation centroids under non-Gaussian distortion by defining a threshold derived from clustering validity metrics, and thus relies on symbol quality evaluation rather than explicit bit-level comparison. In the RoF scenario impaired by PN, the EVM remains an appropriate performance indicator, as it quantifies the geometric deviation of received symbols from their ideal positions and is commonly transformed into BER using well-established analytical relationships reported in the literature [27]. This approach is valid here because clustering-based demodulation directly redefines decision regions based on centroid estimation without altering the underlying constellation geometry; therefore, the EVM continues to reflect symbol quality even under the observed phase noise and nonlinear distortions.

$$BER = \frac{2\left(1-\sqrt{M}\right)}{\log_2(M)} erfc\left(\sqrt{\frac{3\times EVM_{dB}}{2\left(M-1\right)}}\right) \qquad (7)$$

Figure 3b presents the BER performance results of 12+4 PSK modulation format with the same type of lines and markers used for the 16-QAM. The curve for the FCM algorithm shows the worst performance compared with the other methods. Possibly this is due to the non-Gaussian shape of the clusters (ellipsoidal shape), a higher overlap, and hence a lower Euclidean distance among centroids. Similar performance curves are shown for the k-means and conventional demodulation algorithms. However, the GK-FCM algorithm improves the performance by 0.7 dB for the BER threshold of 10-2. The constellation in the inset of Figure 3b is obtained with the FCM-GK algorithm; it has an EVM equal to 20.9% and exhibits a lower separation among clusters in the external ring. However, clustering-based demodulation using GK-FCM performs better in both cases, 16-QAM and 12+4 PSK modulation formats, over 25 dB in the OSNR scale. To evaluate the integrity of the clustering-based demodulation algorithms, we perform estimation for the different validation indices introduced, aided by the Matlab (TM) functions available for the Fuzzy Clustering and Data Analysis Toolbox [26].

In Table I, the results are presented for the validation indexes under the same OSNR levels (16, 26, and 36 dB) using the conventional demodulation technique and the three evaluated clustering demodulation methods. The first three rows show the validation indexes only for three OSNR levels (16dB for high-noise level, 26 dB for medium-noise level, and 36 dB for low-noise level). We can observe that a better classification is performed for the lower-noise level at 36 dB, attaining a PC index of 0.81 using GK-FCM; this index decreases for the other noise levels. Also, the SC index reflects a better partition with the lowest index value for 36 dB and increases with a higher noise level (reduced OSNR). The S index has the lowest distance separation for the low-noise scenario at 36dB, and the other indices' values are proportional to the noise increase.

Table 1. Index validation using clustering-based demodulation for 16-QAM

OSNR (dB)	PC	SC	S	XB	DI	Winner Algorithm	Criteria
36	0.8100	0.2170	1.295E-04	11.11	0.201		
26	0.7210	0.3097	2.031E-04	10.72	0.0049	GK-FCM	The higher value is the best.
16	0.6227	0.4321	2.792E-04	9.81	0.0047		
26	0.5218	0.2863	2.241E-04	9.85	0.0032	FCM	The lower value is the worst.
26	0.7178	0.2987	2.578E-04	10.43	0.0045	k-means	Medium value

Source: own elaboration.

The XB index shows that it minimizes the index value for higher-noise levels (a value of 9.81 for the highest noise OSNR level), and it is consistent because the optimal number of clusters should minimize this index, showing a better separation and compactness for the clusters. The DI for GK-FCM is higher at low noise levels and validates the "*compact and well-separated clusters*" for the whole data symbol received and demodulated. For FCM and k-means, all the indices are consistently compared with the GK-FCM obtained values; the behavior is well represented in the curves of Fig. 3. In Table 2, we show the index values for the 12+4 PSK modulation scheme.

Table 2. Index Validation using clustering-based de-modulation for 12+4 PSK

OSNR (dB)	PC	SC	S	XB	DI	Winner Algorithm	Criteria
36	0.7571	0.2353	1.34E-4	18.03	0.0169		
26	0.6963	0.2886	1.65E-4	16.75	0.0045	GK-FCM	The higher value is the best.
16	0.6592	0.3643	2.09E-4	12.57	4.77E-4		
26	0.6623	0.3116	1.92E-4	14.38	0.0013	FCM	The lower value is the worst.
26	0.6857	0.2812	1.59E-4	16.37	0.0038	k-means	Medium value

Source: own elaboration.

Similarly to the 16-QAM, 0.7571 is the highest level obtained for the PC index, decreasing for higher noise levels. The SC index has a lower value for the lowest noise level, and the index increases with the reduction of the OSNR value. For the S index, the lowest value for an OSNR value of 26 dB, comparing the three algorithms, is 1.34e-04 and corresponds to GK-FCM. The XB index related to the known number of clusters produces a minimum value for the highest noise level of 16 dB in the OSNR axis. Finally, DI has the highest value (0.0169) for the lowest noise level at 36 dB with the

GK-FCM. Despite the lower distance among centroids, the validation indices suggest the algorithm identifies compact and well-separated clusters. During symbol-demodulation, k-means outperforms FCM over 1 dB, and exhibits a higher index of 0.0045 for the DI. From these results, advanced clustering can be used for the identification and characterization of distortion in data symbols. However, if Gaussian behavior is identified, the receiver must be set to perform clustering-based demodulation using k-means or the conventional demodulation algorithm because the fixed grid provides similar error performance as demonstrated. However, if non-Gaussian shapes are observed over the constellation, then a better approach to improve the system performance is to use the FCM-GK method because it improves the noise tolerance by extending the decision boundaries. The observed thresholds, such as PC > 10.7 and DI > 0.015, can be used as indicators to switch from conventional demodulation to clustering-based strategies, offering a dynamic and adaptive receiver approach [29]. These findings are consistent with recent approaches that propose adaptive clustering-based demodulation in nonlinear optical channels [12], [30], [31], validating the effectiveness of fuzzy clustering in scenarios with high phase noise and dispersion.

This study demonstrates that clustering validity indices are not only useful for evaluating demodulation quality but also serve as real-time indicators for switching demodulation strategies. Future optical receivers could benefit from integrating such threshold-based decision mechanisms [32]. To evaluate the presence of nonlinear distortion in the received symbol constellations, two cluster validity indices were analyzed: the Dunn Index (DI) and the Xie-Beni Index (XB). The Dunn Index presents a clear transition behavior. While the values remain relatively low at 16 dB (DI = 0.0047) and 26 dB (DI = 0.0049), an increase is observed at 36 dB (DI = 0.201). This change of slope indicates that the clusters, particularly those associated with the outer symbols of the constellation, become significantly more compact and well-separated as the OSNR improves. In particular, XB 10.7 and DI ≥ 0.015 may serve as empirical indicators that the constellation has transitioned into a more structured, linearly behaving form [31] suitable for traditional demodulation. This suggests that the system moves from a distorted, non-Gaussian symbol distribution to a clearer, Gaussian-like configuration beyond this point. The Xie-Beni index, on the other hand, shows a more gradual trend, with values of 9.81, 10.72, and 11.11 at 16, 26, and 36 dB, respectively. Although this index does not show a sharp inflection, its consistent increase also reflects an improvement in cluster definition, as XB is inversely proportional to clustering quality. Taken together, these results suggest that OSNR ≈ 26 dB can be considered a practical threshold for identifying the onset of significant geometric reorganization in the symbol constellation for RoF systems. Below this threshold, the outer symbols are still affected by overlapping and nonlinear distortions at a BER level of 10^{-2}. Above this threshold, the constellation geometry begins to stabilize, and the clusters become more distinguishable, particularly in terms of compactness and separation.

Finally, to assess the computational efficiency of the clustering algorithms, we analyzed their iteration counts and convergence behavior rather than raw execution time, which is implementation and hardware-dependent. The k-means algorithm required the fewest iterations on average, consistent with its lower per-iteration complexity of $O(nkd)$, where n is the number of samples, k is the number of clusters, and d is the dimensionality. In contrast, FCM and GK-FCM involve additional membership updates and covariance matrix computations, resulting in higher theoretical complexity, approximately $O(nkd)$ per iteration for FCM and $O(nkd^2)$ for GK-FCM. Despite this overhead, GK-FCM demonstrated superior performance in handling ellipsoidal, non-Gaussian cluster shapes, particularly for the 12+4 PSK format, where Euclidean-distance-based k-means degraded more significantly. These observations underscore the trade-off between computational load and improved demodulation accuracy when employing advanced fuzzy clustering methods in distorted constellations. Although covariance matrix computation increases per-iteration complexity and processing time, this cost is balanced by the removal of conventional equalization steps, since IQ imbalance and phase noise are inherently mitigated through the clustering-based demodulation.

CONCLUSIONS

We proposed and experimentally validated advanced clustering-based techniques for distortion identification and symbol demodulation in 16-QAM and 4+12 PSK optical constellations. Among the evaluated methods, the Gustafson-Kessel fuzzy c-means algorithm (GK-FCM) exhibited superior performance in detecting and compensating non-Gaussian distortions, particularly under moderate to high OSNR conditions.

The use of clustering validity metrics enabled the quantification of cluster compactness and separation, offering a practical way to detect nonlinear distortion patterns through threshold analysis. Experimental analysis (Tables 1 and 2) shows that the combined behavior of DI and XB indices establishes a practical threshold for identifying non-Gaussian distortion, with $DI \geq 0.015$ and $XB \geq 10.7$ corresponding to the OSNR transition where clustering-assisted demodulation improves performance. These indices are therefore valuable tools for designing adaptive demodulation strategies in optical receivers under varying OSNR conditions. The obtained results confirmed that these metrics, when used as distortion indicators, support adaptive switching to more adaptive clustering-based demodulation strategies. This decision mechanism led to OSNR gains of up to 2.1 dB for 16-QAM and 0.7 dB for 4+12 PSK at a BER of 10^{-2}, enhancing overall receiver performance without requiring phase-offset or IQ imbalance compensation. These findings demonstrate that clustering validity indices not only improve demodulation performance but can also serve as indicators of nonlinear distortion, supporting the development of intelligent, channel-aware, non-Gaussian noise-tolerant optical systems.

REFERENCES

[1] A. Dogra, R. K. Jha, and S. Jain, "A Survey on Beyond 5G Network With the Advent of 6G: Architecture and Emerging Technologies," *IEEE Access*, vol. 9, pp. 67512–67547, 2021, doi: 10.1109/ACCESS.2020.3031234.

[2] D. Marpaung, "High dynamic range analog photonic links design and simulation," University of Twente, 2021.

[3] J. Liu, X. Wu, C. Huang, H. K. Tsang, and C. Shu, "Compensation of Dispersion-Induced Power Fading in Analog Photonic Links by Gain-Transparent SBS," *IEEE Photonics Technol. Lett.*, vol. 30, no. 8, pp. 688–691, 2018, doi: 10.1109/LPT.2018.2812188.

[4] V. A. Thomas, M. El-Hajjar, and L. Hanzo, "Performance Improvement and Cost Reduction Techniques for Radio Over Fiber Communications," *IEEE Commun. Surv. Tutorials*, vol. 17, no. 2, pp. 627–670, 2015, doi: 10.1109/COMST.2015.2394911.

[5] L. Manoliu, D. Wrana, B. Schoch, S. Haussmann, A. Tessmann, and I. Kallfass, "Frequency and Phase Investigation of the Local Oscillator Offset in a W-Band Satellite Communication Link," in 2023 *53rd European Microwave Conference* (EuMC), 2023, pp. 360–363, doi: 10.23919/EuMC58039.2023.10290372.

[6] L. Li, G. Zhang, X. Zheng, S. Li, H. Zhang, and B. Zhou, "Phase Noise Suppression for Single-Sideband Modulation Radio-Over-Fiber Systems Adopting Optical Spectrum Processing," *IEEE Photonics Technol. Lett.*, vol. 25, no. 11, pp. 1024–1026, 2013, doi: 10.1109/LPT.2013.2258901.

[7] P. Li *et al.*, "Multi-IF-Over-Fiber Based Mobile Fronthaul With Blind Linearization and Flexible Dispersion Induced Bandwidth Penalty Mitigation," *J. Light. Technol.*, vol. 37, no. 4, pp. 1424–1433, 2019, [Online]. Available: http://jlt.osa.org/abstract.cfm?URI=jlt-37-4-1424.

[8] Y. Han, S. Yu, M. Li, J. Yang, and W. Gu, "An SVM-Based Detection for Coherent Optical APSK Systems With Nonlinear Phase Noise," *IEEE Photonics J.*, vol. 6, no. 5, pp. 1–10, 2014, doi: 10.1109/JPHOT.2014.2357424.

[9] Y. Cui *et al.*, "Overcoming Chromatic-Dispersion-Induced Power Fading in ROF Links Employing Parallel Modulators," *IEEE Photonics Technol. Lett.*, vol. 24, no. 14, pp. 1173–1175, 2012, doi: 10.1109/LPT.2012.2192422.

[10] L. Pakala and B. Schmauss, "Two stage extended Kalman filtering for joint compensation of frequency offset, linear and nonlinear phase noise and amplitude noise in coherent QAM systems," in *2017 19th International Conference on Transparent Optical Networks* (ICTON), 2017, pp. 1–4, doi: 10.1109/ICTON.2017.8024972.

[11] D. L. N. S. Inti, "Time-Varying Frequency Selective IQ Imbalance Estimation and Compensation," Virginia Polytechnic Institute and State University, 2017.

[12] M. Solarte-Sanchez, D. Marquez-Viloria, A. E. Castro-Ospina, E. Reyes-Vera, N. Guerrero-Gonzalez, and J. Botero-Valencia, "m-QAM Receiver Based on Data Stream Spectral Clustering for Optical Channels Dominated by Nonlinear Phase Noise," *Algorithms*, vol. 17, no. 12. 2024, doi: 10.3390/a17120553.

[13] D. Wang et al., "KNN-based detector for coherent optical systems in presence of nonlinear phase noise," in *2016 21st OptoElectronics and Communications Conference* (OECC) *held jointly with 2016 International Conference on Photonics in Switching* (PS), 2016, pp. 1–3.

[14] J. J. Granada Torres, S. Varughese, S. E. Ralph, A. M. Cárdenas Soto, and N. G. González, "Clustering in Short Time Windows for Nonsymmetrical Demodulation in 16QAM Overlapped WDM Channels," in *Advanced Photonics 2017* (IPR, NOMA, *Sensors, Networks,* SPP*Com,* PS), 2017, p. SpM3F.2, doi: 10.1364/SPPCOM.2017.SpM3F.2.

[15] A. F. Eduardo, G. T. Jhon James, C. S. Ana María, and G. G.-O. F. T. Neil, "Radio-over-Fiber Signal Demodulation in the Presence of Non-Gaussian Distortions based on Subregion Constellation Processing," *Opt. Fiber Technol.*, vol. 45IS–6, pp. 741–759, 2019, doi: https://doi.org/10.1016/eduardo.yofte.2019.08.004.

[16] D. Lu et al., "100Gb/s PAM-4 VCSEL Driver and TIA for Short-Reach 400G-1.6T Optical Interconnects," in *2021* IEEE *Asia Pacific Conference on Circuit and Systems* (APCCAS), 2021, pp. 253–256, doi: 10.1109/APCCAS51387.2021.9687808.

[17] X. Guan, A. Omidi, M. Zeng, and L. A. Rusch, "Experimental Demonstration of a Constellation Shaped via Deep Learning and Robust to Residual-Phase-Noise," in *Conference on Lasers and Electro-Optics*, 2022, p. SW4E.2, doi: 10.1364/CLEO_SI.2022.SW4E.2.

[18] F. Ali, H. Afsar, A. Alshamrani, and A. Armghan, "Machine learning-based mitigation of thermal and nonlinear impairments in optical communication grids," *Opt. Laser Technol.*, vol. 182, p. 112090, 2025, doi: https://doi.org/10.1016/j.optlastec.2024.112090.

[19] A. Kakkar *et al.*, "Impact of local oscillator frequency noise on coherent optical systems with electronic dispersion compensation," *Opt. Express*, vol. 23, no. 9, pp. 11221–11226, 2015, doi: 10.1364/OE.23.011221.

[20] J. Zhang, M. Gao, W. Chen, and G. Shen, "Non-Data-Aidedk-Nearest Neighbors Technique for Optical Fiber Nonlinearity Mitigation," *J. Light. Technol.*, vol. 36, no. 17, pp. 3564–3572, 2018, doi: 10.1109/JLT.2018.2837689.

[21] J. M. Cebrian, B. Imbernón, J. Soto, and J. M. Cecilia, "Evaluation of Clustering Algorithms on HPC Platforms," *Mathematics*, vol. 9, no. 17. 2021, doi: 10.3390/math9172156.

[22] D. E. Gustafson and W. C. Kessel, "Fuzzy clustering with a fuzzy covariance matrix," in *1978 IEEE Conference on Decision and Control including the 17th Symposium on Adaptive Processes*, 1978, pp. 761–766, doi: 10.1109/CDC.1978.268028.

[23] J. Lasserre, F. Ruiz, and T. Spina, "A double-suppressed possibilistic fuzzy Gustafson–Kessel clustering algorithm (DS-PFGK)," *Knowledge-Based Syst.*, vol. 275, p. 110736, 2023, doi: 10.1016/j.knosys.2023.110736.

[24] D.-W. Kim and K. H. Lee, "A new validity measure for fuzzy c-means clustering." 2024, doi: 10.48550/arXiv.2407.06774.

[25] P. Ghelfi, A. Bogoni, E. Avendaño Fernández, G. Serafino, A. M. Cardenas Soto, and N. Guerrero Gonzalez, "Machine Learning Techniques to Mitigate Nonlinear Phase Noise in Moderate Baud Rate Optical Communication Systems," Y. (Cindy) Yi, Ed. Rijeka: IntechOpen, 2019.

[26] B. Balasko, J. Abonyi, and B. Feil, "MATLAB files for 'Manual for Fuzzy Clustering and Data Analysis Toolbox (For Use with Matlab).'" Jul. 07, 2014.

[27] L. Vendramin, M. C. Naldi, and R. J. G. B. Campello, "Fuzzy Clustering Algorithms and Validity Indices for Distributed Data BT - Partitional Clustering Algorithms," M. E. Celebi, Ed. Cham: Springer International Publishing, 2015, pp. 147–192.

[28] E. A. Fernández, J. J. GranadaTorres, A. M. Cárdenas Soto, and N. G. González, "Geometric Constellation Shaping with Demodulation based-on Clustering to mitigate Phase-Noise in Radio-over-Fiber Systems," in *Latin America Optics and Photonics Conference*, 2018, p. Tu5E.3, doi: 10.1364/LAOP.2018.Tu5E.3.

[29] T. Zhao, A. Nehorai, and B. Porat, "K-means clustering-based data detection and symbol-timing recovery for burst-mode optical receiver," *IEEE Trans. Commun.*, vol. 54, no. 8, pp. 1492–1501, 2006, doi: 10.1109/TCOMM.2006.878840.

[30] D. Wang et al., "Intelligent constellation diagram analyzer using convolutional neural network-based deep learning," *Opt. Express*, vol. 25, no. 15, pp. 17150–17166, 2017, doi: 10.1364/OE.25.017150.

[31] R. T. Jones et al., "Geometric Constellation Shaping for Fiber Optic Communication Systems via End-to-end Learning," Oct. 2018, Accessed: Apr. 16, 2019. [Online]. Available: http://arxiv.org/abs/1810.00774.

[32] M. A. Amirabadi, S. A. Nezamalhosseini, M. H. Kahaei, and L. R. Chen, "A Survey on Machine and Deep Learning for Optical Communications." 2024, [Online]. Available: https://arxiv.org/abs/2412.17826.