Article

Enhancing GNSS PPP Algorithms With AI: Towards Mitigating Multipath Effects

Álvaro Tena [1]*, Adrián Chamorro [2] and Jesús David Calle [3]

[1] GMV; alvaro.tena.tamayo@gmv.com
[2] GMV; achamorro@gmv.com
[3] GMV; jdcalle@gmv.com
* Correspondence: alvaro.tena.tamayo@gmv.com

Abstract: Nowadays, high precision and reliability of Global Navigation Satellite Systems are increasingly important in positioning applications. Machine Learning is used to improve the performance of the GSHARP PPP algorithm by reducing the effect of multipath on GNSS measurements. The clustering analysis is conducted on the primary GNSS data points with the goal of discovering and analyzing patterns in the multipath interference. This study represents an early attempt to apply AI to the GSHARP PPP algorithm. Since Lightweight Machine Learning is used in this research, it is easier to integrate and might serve as a stepping stone to the application of advanced deep learning. About 50 hours of data collected from different environments (e.g. highways and urban areas) serves as the training data for these algorithms which ensures their robustness and real-world applicability. The use of Machine Learning clustering inside the PPP algorithm serves as a way to improve its performance against multipath effects, as well as provide a platform for subsequent development of precision GNSS systems through AI technologies.

Keywords: GNSS; multipath; machine learning; multipath detection; multipath mitigation

Citation: Tena, A.; Chamorro, A.; Calle, D. Enhancing GNSS PPP Algorithms With AI: Towards Mitigating Multipath Effects.

1. Introduction

Today, global navigation satellite systems (GNSS) are an important tool for determining the location and time of any point on earth. Their use in urban and semi-urban environments is constrained due to multipath effects which occur when a satellite signal reflects or diffracts off the buildings surrounding before it reaches the receiver through several paths. This causes a significant impact on code measurement and phase error which poses a great challenge to the accuracy of the GNSS system, especially for applications requiring high precision.

Multipath interference is particularly challenging when signals are received from low elevation satellites affected by nearby structures that cause non-line-of-sight (NLOS) signals that distort the direct or line-of-sight (LOS) signal. These signals induce variations in amplitude, delay, and phase, which can significantly alter the integrity of the received signal [1].

Strategies to mitigate multipath have historically focused on improvements from receiver design and antenna location or type, as discussed extensively in the literature. However, these solutions may be insufficient or impractical in certain contexts. Although recent investigations have expanded the toolkit for addressing multipath effects, incorporating sophisticated data processing techniques directly within receiver systems as the use of different Delay Lock Loops (DLL) estimators, Code-Minus-Carrier (CMC)-based approaches [2] or continuous time-series C/N_0 [3], even a Convolutional Neural Network (CNN) that automatically extracts features and applies a binary classifier on the final layers of the network from raw correlator data in the form of doppler image tensors and code delay offsets [4].

However, there are also recent approaches from the positioning algorithm to mitigate the effects of multipath in GNSS signals, which have served as a basis for this project as will be seen below. In particular, we can observe approaches using CMC only on signals from the Indian satellite constellation, NavIC [5], and other approaches using Code Rate Consistency (CRC), a term that we will explain in depth later on, together with satellite elevation, C/N_0 and pseudorange residuals [6].

2. Materials and Methods

GPS measurements from L1C, L2C and L5Q signals, Galileo measurements from E1C, E5aQ and E5bQ signals and BeiDou measurements from B1C and B2Ap signals have been used in this study. They were collected, mainly in kinematic conditions, in the surroundings and the center of Madrid, Spain. These data belong to several data campaigns from different projects from February 2023 to January 2024 and they have been reused for the analysis presented in this study. The data available is about 50 hours in total. Due to the complexity of GNSS signals and environments, a single feature is not enough to classify an observation. Thus, a set of features is needed to appropriately represent the multipath. To process this data, we utilize GSharp PPP, a PPP algorithm designed by GMV [7]. When these data are processed through GSharp PPP we obtain a combination of different features, such as pseudorange, phase, Doppler shift, signal-to-noise ratio (SNR), elevation, pseudorange residuals and ionospheric delay calculated from the previous epoch. With all these features we can set the base parameters for any satellite of any constellation and frequency.

2.1. Feature Engineering

A good selection of features and removal of irrelevant or redundant ones in machine learning (ML) is key to handling high-dimensional data and incorporating domain knowledge can help with that, reducing computation time and model complexity improving learning accuracy, and facilitating a better understanding of the model or data [8,9]. Therefore, we need to conscientiously select what features will improve the clustering accuracy.

1. SNR: It is an important metric that helps determine the quality of a received signal against the background noise level. The original meaning of this term is the ratio of the power of the signal versus the power of the noise influencing signal clarity and integrity. The SNR values higher than a certain threshold are often associated with having a clear signal with little interference and this is very important for successful data transmission and receiving. In high SNR environments, it is easier for GNSS receivers to generate accurate observables from the tracked signals. Nonetheless, during nominally fluctuating noise conditions or in the presence of other signals that can interfere or create noise modification, the signal can be masked or falsely magnified; implying that the quality of the observables will evolve together with the C/N0 in a similar way by destructive and constructive multipath.
2. Elevation Angle: It is a traditional and effective way to set up weighting coefficients in the Processing SWs to diminish the role of multipath and NLOS signal reception in positioning. As an example, satellite signals which are located at a higher elevation angle usually have a smaller chance to be blocked and reflected by buildings. However, this cannot be fully generalized. Because of the width and height of buildings in urban centers, satellite signals at high elevation angles may be NLOS signals, while signals at low elevation angles may indeed be direct signals.
3. Code Residuals: The code residuals of a least square PVT solution where position, clock bias, inter-system biases and inter-frequency biases are estimated (in snapshot mode) has been used also as a feature. Checks to detect residual outliers is also a traditional method for trying to discard misleading measurements, although they show limited performance for that goal under conditions with low number of satellites.

Despite this fact, this feature was considered interesting to be included in the analysis and training.

4. Ionospheric Delay: GNSS signals are delayed as they travel through the ionosphere, a part of Earth's atmosphere ionized by solar radiation. This delay varies with the electron density, which changes with solar activity and the time of day. The ionosphere is dispersive, causing delays that differ by signal frequency. As we do not know the current value of the Ionospheric delay we use the value of the previous epoch calculated by the GSharp algorithm, which provides an accurate ionospheric delay observation in near-real time because it receives ionospheric corrections from the GSharp Correction Service. The delay does not change as fast as the algorithm calculates a position, 10 Hz, so we can ensure low error the estimation due to the fact of using ionospheric delay from the previous epoch.

5. ΔCMC: A common approach to isolate multipath is the CMC metric (Carrier Minus Code). Due to the common terms in the pseudorange and phase equations, they will cancel each other when subtracted with the exception of the ionosphere delay (and carrier phase ambiguity), which for phase measurements suffers negative delay and for code measurements positive delay, being modeled for a satellite s, a frequency i, and an epoch k as:

$$CMC^s_{i,k} = \rho^s_{i,k} - \phi^s_{i,k} = 2I^s_{i,k} + N^s_{i,k}\lambda_i + MP^s_{i,k} - mp^s_{i,k} + \epsilon^s_{i,k} - \zeta^s_{i,k}, \tag{1}$$

where $I^s_{i,k}$ is the ionospheric delay in meters, $N^s_{i,k}$ and λ_i are the integer ambiguity expressed in cycles and the wavelength of frequency i, leaving the remaining terms ($MP^s_{i,k}$, $\epsilon^s_{i,k}$ and $mp^s_{i,k}$, $\zeta^s_{i,k}$) representing the multipath and thermal noise errors in code and phase measurements, all in meters.
Since we have the ionospheric delay, we could simply subtract twice the delay from 1, but as seen in Caamano et al., that new feature known as CMC_{Dfree} has a linear combination of the phase ambiguities from the two frequencies. Therefore, it is preferred to use the rate of change of the CMC, which means that the ionospheric delay is not completely eliminated and that you accumulate twice the change of the ionospheric delay ($2\dot{I}$) besides the rate of noise and multipath ($\dot{\epsilon}$ and $\dot{MP}$). But as they remark, the ionospheric rate component is negligible in nominal conditions. The ΔCMC equation is expressed as follows:

$$\Delta CMC_k = \frac{1}{\Delta t}(CMC_k - CMC_{k-1}) \approx 2\dot{I}_k + \dot{MP}_k + \dot{\epsilon}_k, \tag{2}$$

and, in addition, we propose the alternative of making use of the absolute value of ΔCMC to more effectively highlight significant fluctuations and mitigate the impact of directional changes in the ionospheric delay and other error components, thus providing a clearer indication of anomalies in GNSS signal integrity.

6. CRC: As seen in Wang et al., the interference of multipath/NLOS signal on the frequency control loop is less than that on the code tracking loop, the consistency between the pseudorange change rate and the Doppler frequency shift can reflect the degree of reflective signal interference. It can be expressed as:

$$CRC = |\Delta\rho - \dot{\rho} \cdot \Delta t|, \tag{3}$$

where Δp stands for the pseudorange variation and Δt for the time interval. As per Doppler effect, the pseudorange rate $\dot{p}$ is calculated from the Doppler shift term:

$$\dot{\rho} = -\lambda_i \cdot f_{D_i}, \tag{4}$$

where the wavelength of frequency i is represented as λ_i and the Doppler shift is f_{D_i} in Hz.

The rest of the features can be omitted for the reasons explained above, because it contains irrelevant or redundant data.

2.2. *K*-Means Clustering Algorithm

The *K*-Means algorithm groups the data by attempting to separate the samples into n groups of equal variance, minimizing a criterion known as the sum of inertia or sum of squares within a cluster, which can be modeled as follows:

$$\sum_{i=0}^{n} \min_{\mu_j \in C}(\|x_i - \mu_j\|^2), \tag{5}$$

where n is the number of samples, μ_j is the mean of the samples of a cluster C and x_i is the sample.

The *K*-Means algorithm needs, previously to start with, the number of clusters k. Although we could differentiate 2 or 3 types of GNSS signals in reference to multipath interference (LOS, multipath and/or NLOS), that might not be the best choice to establish the number of clusters, so we will need to set some experiments trying different values for k. In addition, we will have to perform an hyperparameters optimization to find the most suitable ones to increase the accuracy.

To evaluate the resulting *K*-Means model, we propose the use of 3 metrics widely used in unsupervised learning, such as the Silhouette Coefficient, the Davies–Bouldin index (DBI) and the Calinski–Harabasz index (CHI).

- Silhouette Coefficient: It evaluates the quality of clustering based on the cohesion and separation of clusters. It is calculated for each point and measures its affinity to its own cluster and to the nearest neighboring clusters. A value close to 1 indicates a high affinity, while a value close to -1 indicates that the point could be better assigned to another cluster. This metric is calculated based on the following formula:

$$S = \frac{b - a}{\max(a, b)}, \tag{6}$$

  where a represents the average intracluster distance and b represents the average distance to the nearest cluster.

- DBI: It is based on the distance between centroids and the dispersion within clusters. A lower value of the metric indicates a better quality of the clusters, with more compact and well-separated groups. It is calculated based on the following formulas:

$$DBI = \frac{1}{k}\sum_{i=1}^{k} \max_{i \neq j} R_{ij}, \tag{7}$$

$$R_{ij} = \frac{s_i + s_j}{d_{ij}}, \tag{8}$$

  where k is the number of clusters, s is the average distance between each point and the centroid and d_{ij} is the distance between centroid i and j.

- CHI: Also known as the variance ratio criterion, it evaluates the quality of clustering by comparing the dispersion between clusters with the dispersion within each cluster. A higher value of this index indicates greater separation between clusters and lower dispersion within clusters, which is associated with better clustering quality. It can be calculated using the following formulas:

$$CHI = \frac{BCSS/(k-1)}{WCSS/(n-k)}, \tag{9}$$

$$BCSS = \sum_{i=1}^{k} n_i \|c_i - c\|^2, \tag{10}$$

$$WCSS = \sum_{i=1}^{k} \sum_{x \in C_i} \|x - c_i\|^2, \tag{11}$$

where n_i is the number of points in cluster C_i, c_i is the centroid of C_i, and c is the overall centroid of the data.

3. Results

3.1. Analysis of K-Means Clustering for Multipath Detection

As indicated before, it is not that simple to define the optimal number of clusters as basing one's knowledge on domain knowledge. Hence, the set of experiments for the different number of clusters, 2 to 6, were created. This section will therefore compare the experiments to make sure that the right number of clusters can detect multipath. Having discussed about how K-means algorithm works with different number of clusters, we will look into the performances and see the implications of each of different experimental setup.

3.1.1. Data Analysis

The data used for the experiments correspond to 37 GNSS scenarios in different types of environments, from highway to deep urban, and have been recorded for GPS, GAL and BDS constellations. Because of different complex environments the data needs a previous analysis, to determine the amount of outliers we can preserve or the scalers to be used before training the model.

Figure 1 illustrates the box plots and density functions of those features which are generated as the model. The figures depict the distribution of typical and the most severe values for every single parameter. These outliers are seen mainly in a number of features, in particular delta CMC, and CRC. These inconsistencies prevent us from showing the entire distribution in the delta CMC and CRC graphs because they are noticeable in both graphs.

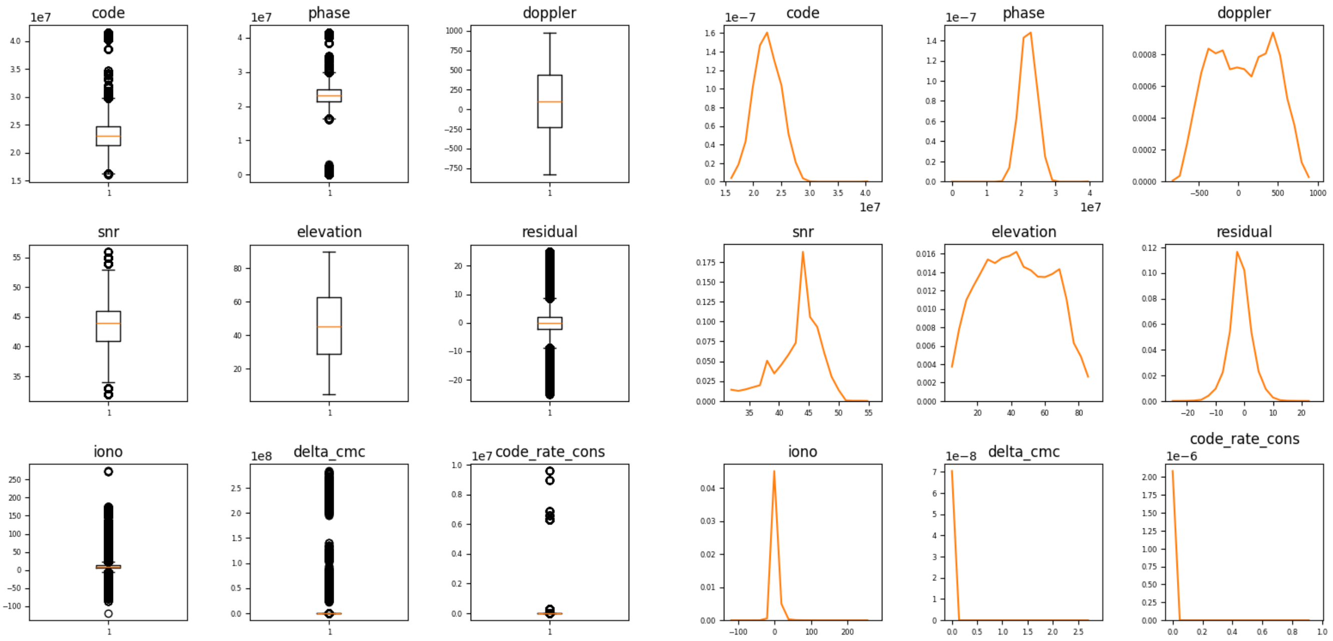

Figure 1. Box plot and Density plot. (**a**) Box plots illustrate the variability and spread of features. (**b**) Density plots highlight the frequency distribution for each feature.

Although the model has no knowledge of which satellite ID or constellation a given observation belongs to, it is important to analyze how the measurements are distributed among the different constellations and IDs, in order to avoid an imbalance that could lead to discrimination of certain observations based on which constellation or satellite ID they come from. In Figure 2 we can see how, although the number of observations does not coincide between IDs or between constellations, there is not a large imbalance between features for which we need to take balancing measures, such as under-sampling or over-sampling.

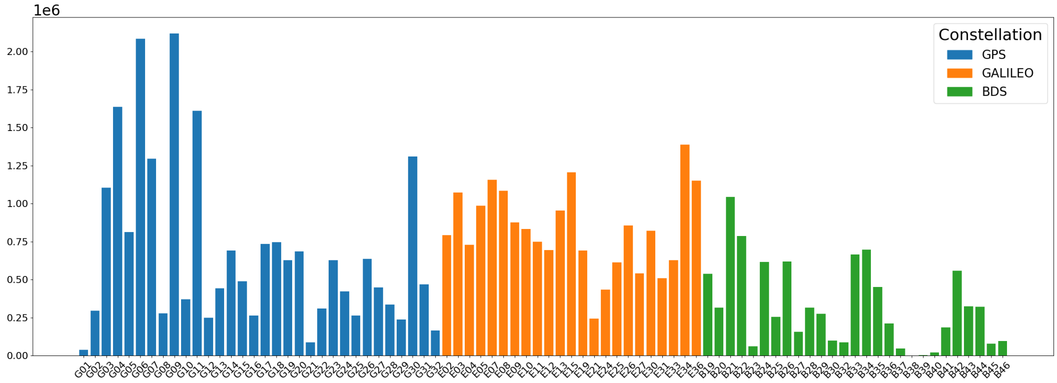

Figure 2. Satellite visibility distribution across GPS (blue), Galileo (orange), and BDS (green) constellations. This histogram illustrates the number of observations per satellite, highlighting the variability in data availability among different satellites within each constellation.

The next step in data processing is to normalize or scale the data to ensure that the different features are not overdimensioned due to data scaling. For this purpose, the following have been applied: standard scalers, scalers that are resistant to outliers (if extreme values are not desired), or scalers that make ratio of values within the range between minimum and maximum, maintaining the distance between all values regardless of their extreme.

3.1.2. Cluster Analysis

To determine the optimal number of clusters (k), we analyzed various metrics across different clustering configurations as shown in table 1 for clustering evaluation, the CHI, DBI, and Silhouette Coefficient.

The CHI metric, a key factor assessing the integrity of clusters internally, is bounded and also quite important for optimizing hyperparameters, however, it lacks a maximum value, so it is not good at comparing experiments. It displays general characteristics but particular conclusions should be made within the context of each experiment, not based on general assumption.

On the contrary, DBI gives a measurement of the quality of clustering directly with the lowest value representing best clustering distribution. Here, the 5-cluster model, possessing the lowest DBI value (0.836727) is recommended as the best model to distinguish the different models compared.

Along with this, the Silhouette Coefficient of the 5 cluster model is higher than the 6 clusters and 4 clusters, thus indicating better cooperation within the same clusters and separation of others. This item echoes the DBI findings, thus supporting the uptake of the 5-cluster model which will help in clearly and distinctly clustering the data by choice.

Table 1. Model metrics for each experiment.

Number of Clusters	CHI	DBI	Silhouette Coefficient
2 Clusters	58711400	0.999	0.457
3 Clusters	92425300	0.917	0.380
4 Clusters	93903700	0.963	0.357
5 Clusters	90250400	0.836	0.368
6 Clusters	83099200	0.925	0.325

3.1.3. Model Analysis

As we have selected k as 5 based on the DBI metric and because of domain knowledge we separate the signal types into 2 or 3 (LOS, multipath and/or NLOS), we understand the results as those 3 signal types but with other additional features or different degrees of multipath affection up to NLOS.

Based on the complete dataset of observations for the different scenarios, we can analyze in a general way the performance of the model. So, if we have a look at Figure 3 we can see the distribution of all these measurements once they have been classified along the entire dataset.

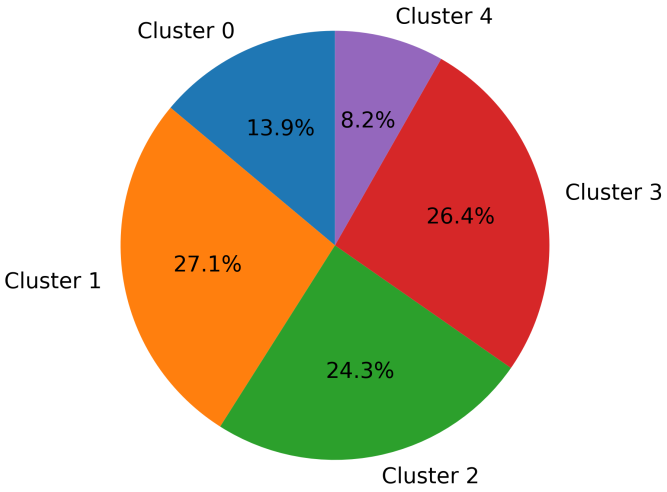

Figure 3. Cluster distribution.

We can further assess the quality of the clusters by examining the relationships between key features of the dataset. A pairwise comparison between features highlights the degree of separation achieved by the clustering process and provides insights into the internal structure of the data, helping validate whether the clusters are meaningful in terms of physical interpretation and contributing to the model's explainability.

Figure 4 illustrates the relationships between the different features and their corresponding clusters, with the diagonal showing the distribution of each feature by cluster, and the off-diagonal plots representing pairwise comparisons. The matrix is symmetric, meaning the lower and upper halves contain the same information with reversed axes, making it easier to observe the segregation performed by the model, as supported by the DBI metric discussed earlier.

Focusing on the elevation-SNR relationship, a traditional metric in GNSS signal analysis, we infer that Cluster 3 groups the cleaner LOS observations, with high elevation and SNR values, indicating minimal multipath interference. Clusters 1 and 2, although at lower elevations, maintain high SNR values, suggesting relatively clean signals with some level of degradation, likely due to lower-elevation LOS with minor multipath interference.

In contrast, Clusters 0 and 4 exhibit the lowest SNR values, representing measurements under significant multipath or NLOS conditions. These clusters likely reflect signals heavily influenced by reflections or obstructions, common in urban or obstructed environments. This clear distinction between clusters based on SNR and elevation is crucial for interpreting signal quality, as it allows us to classify measurements effectively for further use in positioning algorithms.

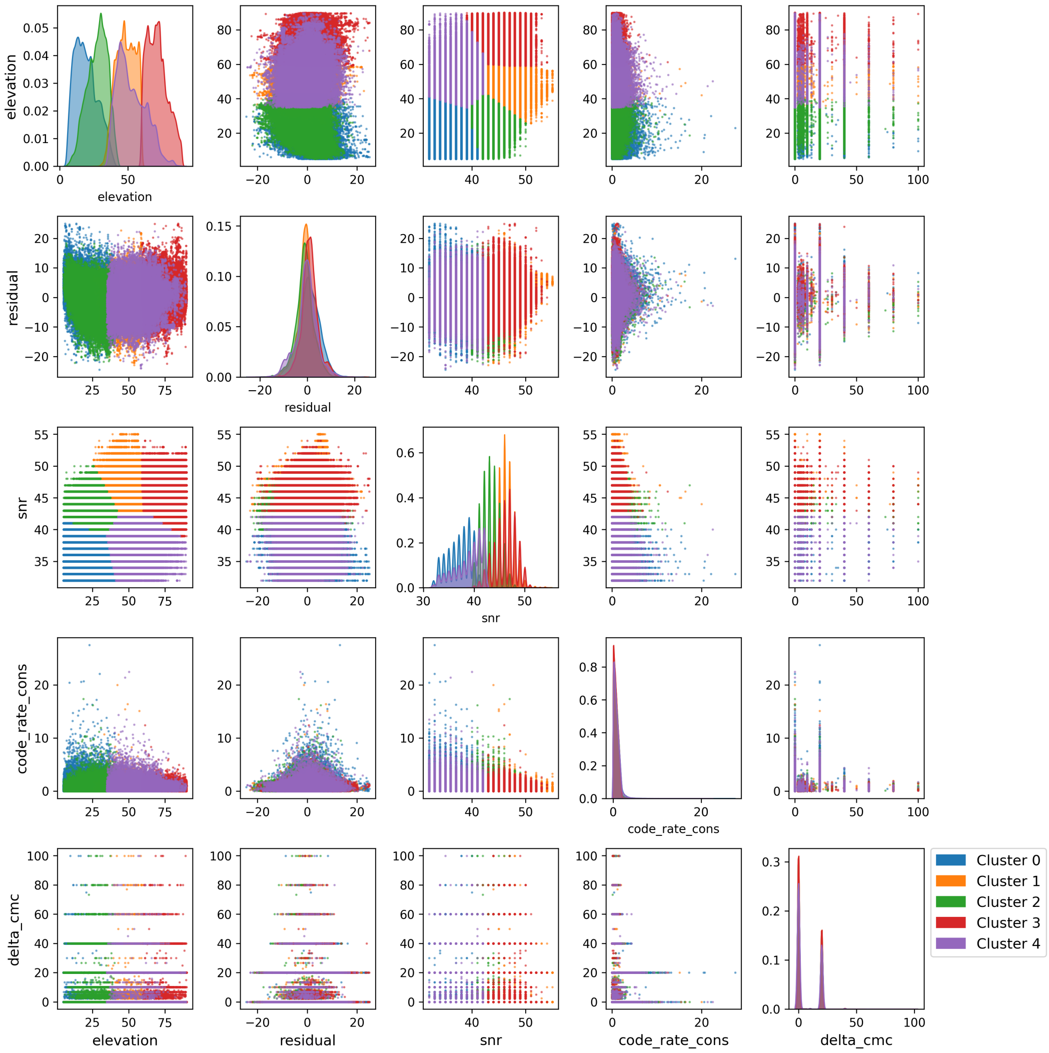

Figure 4. Visualization of multiple features across different GNSS data relationships, grouped by cluster.

3.2. Scenario Analysis with Clusters Data

As a final result, we analyze a scenario with the KMeans output of 5 clusters, covering conditions from open sky to urban environments. In Figure 5, we observe the changing scenario conditions based on the model interpretation. In the first static phase, clusters 0 (blue) and 4 (purple) show fewer measurements, as expected in a semi-urban environment where GNSS signals are relatively strong due to favorable satellite geometry, minimizing multipath and interference under LOS conditions.

Next, in the cyan phase, which combines highway and semi-urban environments, the number of measurements varies across clusters due to changing obstructions in the environment. Highways provide better signal reception than semi-urban areas, causing fluctuations in cluster distribution. In the following navy phase, representing an urban area with wide streets, Cluster 3 (red) remains consistent, indicating better reception from high-elevation satellites, where obstacles have less impact on signal quality.

In contrast, the gray phase, representing a deep urban environment with narrow streets, shows an increase in Cluster 4 (purple), reflecting signals affected by multipath, while Cluster 1 (orange), associated with medium-elevation satellites, remains stable. Two subzones within this urban area, marked in teal (a large avenue) and olive (parks and low buildings), provide more open environments, improving signal reception, as seen in the rise of Cluster 2 (green), where satellites with good SNR and lower elevation dominate due to fewer obstructions.

Returning to the gray deep urban environment, narrow streets reduce the number of Cluster 1 (orange) measurements, indicating further occlusion. In the final phases (navy and cyan), signal reception improves again, reflecting more measurements from high-elevation satellites with strong SNR. Despite similar environmental conditions at different points, there are notable variations in the number of measurements within the same cluster, attributable to changes in both satellite and environment geometry. This highlights that GNSS signal interpretation relies not only on the type of environment but also on satellite positioning and availability at specific times.

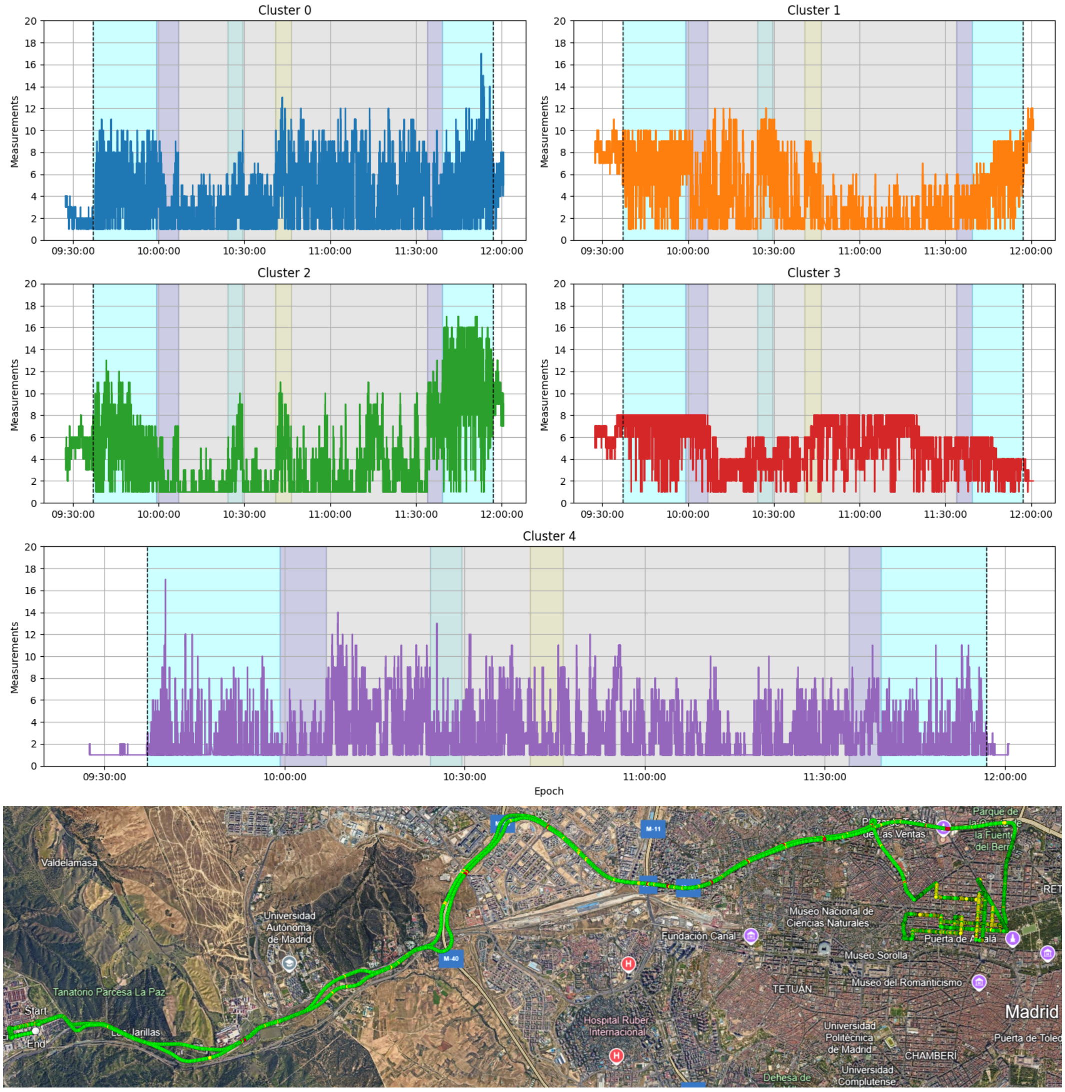

Figure 5. (**a**) KML reference positioning (**b**) Visualization of the variation of cluster assignment to measurements over a GNSS scenario.

4. Discussion

The results demonstrate how the use of AI in GNSS can be beneficial, in this case by helping in the detection of multipath problem in complex environments. Satisfactory clustering of observations at the positioning algorithm level is observed, with a correlation consistent with the expectations. The approach makes this system an agnostic solution to the GNSS receiver in use, making it independent of its model or its quality, although some extra training could be needed.

By applying the model to a particular scenario we can appreciate the nature of that scenario with more detailed information, making it possible for future systems to make use of this information as an additional input for calculating an accurate position. In this paper we have tried to go beyond the analysis and improve the position of the PPP based on this clustering. The approach chosen has been to modify the weights that are given to each measurement before entering the extended Kalman filter (EKF) of GSharp PPP. This process is outside the scope of this paper and the further experimentation will be carried out with the goal to mitigate the impact of measurements with multipath in the positioning obtained by the GSharp PPP.

However, despite the advances presented in this study on understanding and detecting the multipath effect in GNSS signals, complete mitigation remains a complex challenge. Further research is required to determine the most effective mitigation strategies. This paper opens the way for more complex AI models, such as Deep Learning (DL) or Reinforcement Learnign (RL), in order to correctly and accurately adjust the weights of the measurements, avoiding the error that we can induce with our interpretation.

Conflicts of Interest: The authors declare no conflicts of interest.

Abbreviations

The following abbreviations are used in this manuscript:

AI	Artificial Intelligence
CHI	Calinski–Harabasz Index
CMC	Code-Minus-Carrier
CNN	Convolutional Neural Network
CRC	Code Rate Consistency
DBI	Davies–Bouldin Index
DL	Deep Learning
DLL	Delay Lock Loop
DR	Dead Reckoning
EKF	Extended Kalman Filter
GNSS	Global Navigation Satellite Systems
LOS	Line-Of-Sight
ML	Machine Learning
NLOS	Non Line Of Sight
PPP	Precise Point Positioning
RL	Reinforcement Learning
SNR	Signal-To-Noise Ratio

References

1. Braasch, M. Multipath. *Springer Handbook Of Global Navigation Satellite Systems*. pp. 443-468 (2017), https://doi.org/10.1007/978-3-319-42928-1_15
2. Caamano, M., Crespillo, O., Gerbeth, D. & Grosch, A. Detection of GNSS Multipath with Time-Differenced Code-Minus-Carrier for Land-Based Applications. *2020 European Navigation Conference (ENC)*. pp. 1-12 (2020)
3. Kubo, N., Kobayashi, K. & Furukawa, R. GNSS Multipath Detection Using Continuous Time-Series C/N0. *Sensors*. **20** (2020), https://www.mdpi.com/1424-8220/20/14/4059
4. Munin, E., Blais, A. & Couellan, N. Convolutional Neural Network for Multipath Detection in GNSS Receivers. (2019)
5. Shukla, A. & Sinha, S. Unsupervised machine learning approach for multipath classification of NavIC signals. *ION GNSS+, The International Technical Meeting Of The Satellite Division Of The Institute Of Navigation*. (2022)
6. Wang, H., Pan, S., Gao, W., Xia, Y. & Ma, C. Multipath/NLOS Detection Based on K-Means Clustering for GNSS/INS Tightly Coupled System in Urban Areas. *Micromachines*. **13** (2022), https://www.mdpi.com/2072-666X/13/7/1128
7. GMV GSharp PPP. (2023), https://www.gmv.com/en/products/space/gmv-gsharp [Accessed: (accessed on 10 May 2024)]
8. Cai, J., Luo, J., Wang, S. & Yang, S. Feature selection in machine learning: A new perspective. *Neurocomputing*. **300** pp. 70-79 (2018)
9. Roe, K., Jawa, V., Zhang, X., Chute, C., Epstein, J., Matelsky, J., Shpitser, I. & Taylor, C. Feature engineering with clinical expert knowledge: A case study assessment of machine learning model complexity and performance. *PLoS ONE*. **15** (2020)