

Data-Centric Preprocessing for GNSS-Based Orbit Determination via AI, Signal Processing, and Orbital Dynamics

Marta Guimarães^{*1,2,3}, Robert Arthur¹, Cláudia Soares², Chiara Manfretti^{1,3}, Gerarda De Pasquale⁴

¹Neuraspace, Coimbra, Portugal

²NOVA School of Science and Technology (NOVA FCT), Caparica, Portugal

³Technical University of Munich (TUM), Munich, Germany

⁴ESA - ESTEC, Noordwijk, Netherlands

GNSS-based orbit determination relies on input measurements whose quality can vary significantly across mission phases and operating conditions. In such settings, inappropriate data conditioning may either improve estimation robustness or, if misapplied, degrade convergence and numerical stability. This work proposes a data-centric preprocessing pipeline designed to operate safely under heterogeneous data quality conditions. The method combines lightweight physical sanity checks with a structure-aware ML classifier that gates increasingly assumption-dependent signal-processing-based cleaning stages. The pipeline targets localised measurement irregularities across multiple orbital representations while explicitly avoiding over-cleaning in weakly structured segments. In representative scenarios, it agrees with a baseline batch least-squares estimator on obvious outliers and additionally identifies dynamically inconsistent samples that remain invisible to position-only residual analysis. The results show that data-centric AI preprocessing reduces residual dispersion and improves convergence behaviour, resulting in a more stable and statistically consistent orbit determination without modifying the underlying dynamical model.

1 Introduction

This work is motivated by high-stakes operational [Global Navigation Satellite System \(GNSS\)](#)-based [orbit determination \(OD\)](#) use cases, where reliable [OD](#) is a critical enabler for downstream tasks such as collision avoidance, and early mission assessment. In sensitive phases such as launch and early orbit, degraded or inconsistent data can significantly reduce estimation quality or lead to non-convergence of otherwise standard [OD](#) procedures. [GNSS](#)-derived states are noisy, heterogeneous, and affected by multiple interacting error sources. Such effects are persistent rather than exceptional in real settings. The objective here is therefore not to modify or replace established [OD](#) algorithms, but to improve data robustness through a dedicated data-centric preprocessing stage. This perspective aligns with the recent data-centric AI paradigm prioritising dataset quality and cleaning over model changes [1–3].

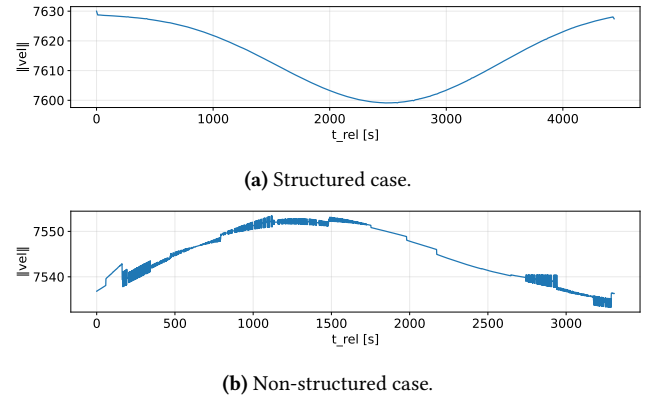


Figure 1: Evolution of the velocity magnitude for a structured and a non-structured case. In the non-structured case, the ambiguous and irregular signal evolution makes simplified dynamical assumptions unreliable.

[Machine learning \(ML\)](#) has been explored in several problems adjacent to the one addressed in this work. In particular, two lines of research are most closely related: (i) the detection of specific measurement degradation sources at the signal or receiver level [4, 5], and (ii) learning-based approaches to orbit estimation, proposed as alternatives or complements to classical [OD](#) techniques [6]. As noted by Siemuri *et al.* [4], however, the application of [ML](#) to [GNSS](#)-based satellite [OD](#) remains comparatively limited.

We adopt the premise that corruption in [GNSS](#) measurements is unavoidable in realistic scenarios. The objective is not error diagnosis nor orbit estimation via [ML](#), but the identification and controlled handling of data segments likely to degrade the statistical consistency of downstream [OD](#). To the best of our knowledge, this pipeline-level formulation has not been explicitly addressed. This formulation is further motivated by the use of real operational data, which introduces challenges beyond controlled benchmark settings. Measurements originate from heterogeneous sources and processing chains, often with limited metadata regarding sensor characteristics or upstream estimation procedures. In particular, the provenance of derived quantities

^{*}Corresponding author. E-Mail: marta.guimaraes@neuraspace.com

such as velocity may be unclear. Rather than explicitly modelling these differences, we pursue a source-agnostic, data-centric approach that favours robust mechanisms compatible with existing OD workflows.

Given these considerations, we propose a data-centric preprocessing pipeline for operational OD that integrates orbital-dynamics filtering, AI-based structure assessment, and signal-processing-driven anomaly removal. The method operates in three stages: (i) base filtering, (ii) structure assessment via ML, and (iii) assumption-controlled anomaly removal, ensuring that increasingly aggressive filters are applied only when supported by the data.

2 Base Filters

The pipeline begins with lightweight sanity and domain filters to remove measurements incompatible with physically meaningful orbital motion.

Segments with fewer than 100 samples or spanning less than 100 s are discarded due to insufficient temporal context.

Grossly unphysical samples are rejected using magnitude thresholds of $\|\text{pos}\| > 10^7$ m and $\|\text{vel}\| > 10^{10}$ m/s.

Finally, skew- and quantile-based filtering is applied to position and velocity magnitudes to suppress heavy tails and reduce the influence of extreme values on subsequent stages.

3 Structure Classification

The goal is to distinguish segments suitable for assumption-based cleaning from those where such assumptions are unreliable. This is formulated as a binary classification between non-structured and structured segments (Figure 1).

Model Choice and Architecture. The data impose several constraints on model selection: the input time series are multivariate, irregularly sampled, unaligned across samples, and of variable length. Moreover, the objective is not temporal localisation, but the global characterisation of structure within a segment. So, convolution-based approaches are a natural choice [7].

We adopt a [convolutional neural network \(CNN\)](#) classifier with a deliberately simple architecture. The model consists of a small number of 1D convolutional layers, followed by average pooling and a binary classification head. Input features include position and velocity magnitudes, as well as Keplerian and equinoctial orbital elements. The classification task is inherently imbalanced, with non-structured cases representing approximately 8% of the data. No explicit class rebalancing strategy was applied, as the objective was to assess the model’s baseline behaviour. Stratified splits were used across training, validation, and test sets to preserve class proportions.

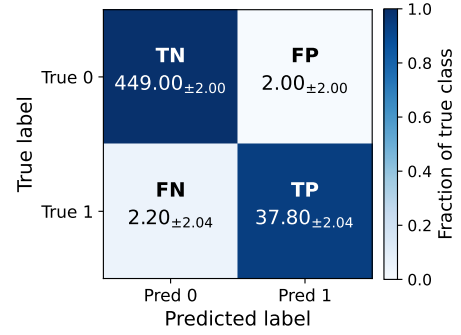


Figure 2: Confusion matrix over 10 independent trials. Values indicate the $\text{mean} \pm \text{std}$ of sample counts in each category. TP: true positives; TN: true negatives; FP: false positives; FN: false negatives.

Classification Results. The classifier was evaluated over 10 independent trials using stratified splits. Table 1 reports the mean and standard deviation of precision, recall, and F1-score [8]. Performance is consistent across runs despite class imbalance, with an average F1-score of 0.947.

The confusion matrix (Figure 2) shows balanced false positive and false negative rates, with a small absolute number of false negatives. Operationally, false negatives (non-structured segments classified as structured) are more critical, as they may trigger assumption-dependent cleaning where its premises do not hold. False positives are conservative: structured segments bypass cleaning and proceed to OD with relaxed settings, affecting efficiency but not estimation integrity.

Non-structured segments often coincide with cases where OD fails to converge or remains ambiguous, even under expert inspection. Automatically flagging such segments therefore supports safer downstream decisions, including specialised processing or exclusion. A binary formulation reflects this operational objective. More granular multi-class approaches could offer finer guidance but would exacerbate imbalance and labelling challenges; such extensions are left for future work.

Precision	Recall	F1-Score
0.952 ± 0.045	0.945 ± 0.051	0.947 ± 0.036

Table 1: $\text{mean} \pm \text{std}$ values over 10 independent trials.

4 Anomaly Cleaning

This section describes the staged cleaning pipeline applied to segments classified as structured. The stages operate under progressively stronger assumptions, enabling conservative behaviour in weakly structured regimes while allowing more aggressive, assumption-controlled correc-

tions when supported by the data.

4.1 Stages

Dynamical Content Index. For each segment j , a *dynamical content index* d_j quantifies orbital variability over the sample set $\mathcal{S}_j$:

$$d_j = \text{IQR}(\|\mathbf{v}\|) + \text{IQR}(\text{sma}) + \text{IQR}(\text{ecc}) + \text{IQR}(\text{inc}), \quad (1)$$

where $\text{IQR}(\cdot)$ is the interquartile range computed over $\mathcal{S}_j$. The index measures the spread of physically meaningful orbital quantities, yielding larger values for segments with clear dynamical variation and smaller values for nearly flat behaviour. The use of interquartile ranges ensures robustness to outliers and scale invariance across regimes. Based on d_j , segments are assigned to one of three dynamical regimes that control cleaning aggressiveness: *flat* (no additional detection beyond base filters), *mid* (peak-based detection with relaxed thresholds), and *rich* (full anomaly cleaning with default parameters).

Edge Cleaning. Boundary instabilities are common in OD and difficult to detect with generic anomaly detectors due to limited temporal context. A dedicated edge-cleaning stage suppresses unstable behaviour at segment boundaries while preserving the interior. The method uses a robust anomaly score based on first-order temporal differences, $\Delta X_{t,f}$, of Keplerian and equinoctial elements. For each feature f ,

$$Z_{t,f} = \frac{\Delta X_{t,f} - \text{median}(\Delta X_{\cdot,f})}{\text{MAD}(\Delta X_{\cdot,f})}, \quad (2)$$

with per-sample score $s(t) = \sum_f |Z_{t,f}|$. Edge instability is assessed by comparing score statistics in boundary windows with those of the interior. The deviation is quantified using

$$J = \frac{\text{median}(s_{\text{edge}}) - \text{median}(s_{\text{int}})}{\text{MAD}(s_{\text{int}})}. \quad (3)$$

If the deviation exceeds a predefined threshold, samples are iteratively trimmed from the corresponding boundary until the score distribution becomes consistent with the interior reference or a maximum trimming fraction is reached.

Peak Detection. This stage targets localised deviations within the segment interior. It uses the same robust normalisation as edge cleaning, applied directly to feature values:

$$Z_{t,f} = \frac{X_{t,f} - \text{median}(X_{\cdot,f})}{\text{MAD}(X_{\cdot,f})}, \quad (4)$$

with anomaly score $s(t) = \sum_f |Z_{t,f}|$. Candidate anomalies are identified as local extrema of the score signal $s(t)$, considering both peaks and dips. Each extremum is associated with a prominence value that quantifies its contrast relative to the surrounding neighbourhood.

Because neighbouring extrema often correspond to the same underlying excursion, candidates occurring in adjacent samples are grouped into short clusters. For each cluster, a single representative is selected using a deterministic ranking: (i) highest prominence; (ii) largest absolute deviation from a five-sample local median if prominence is unavailable or tied; (iii) earliest in time if still tied.

Neighbour Expansion. A one-dimensional neighbour expansion step is applied to reduce fragmented detections. Depending on the selected mode, it can (i) fill single-sample gaps between adjacent detections (*close*), (ii) expand detections to immediately neighbouring samples (*dilate*), or (iii) apply both. This step operates solely on the detection mask and introduces no additional assumptions about the underlying signal.

4.2 Cleaning Results

Figure 3 shows the cleaning process applied to a representative data segment. Panels 1 & 2 show raw position and velocity magnitudes with samples rejected by domain- and skew-based filters. Panels 3 & 4 display the corresponding cleaned magnitudes and components. Panels 5 & 6 show the resulting Keplerian and equinoctial elements. Panel 7 presents the anomaly score with detections from the peak, edge, and neighbour stages. Vertical yellow shading indicates detected peaks, while green bars denote measurements rejected by a baseline batch least-squares OD [9].

This example compares the proposed pipeline with the OD baseline. Both methods agree on out-of-domain samples, which are obvious measurement failures and trivially rejected by both. Beyond these cases, clear differences emerge. Since the OD baseline relies on position-only residuals, it cannot detect anomalies that arise from velocity inconsistencies or higher-order feature distortions. In contrast, the skew-based filter captures many such cases, particularly visible in the velocity magnitude panel, where samples that clearly break the orbital trend remain unflagged by OD. This highlights a key advantage of the proposed approach: it detects irregularities invisible to a purely position-based estimator.

After domain and skew filtering, the position and velocity signals become coherent, and the resulting orbital elements exhibit regular behaviour with only a few local distortions aligned with score peaks. This indicates that the detected anomalies correspond to meaningful feature-level inconsistencies rather than noise.

While Figure 3 provides qualitative evidence, Table 2 quantifies the impact of the cumulative filtering stages on OD performance within the same window¹. As expected,

¹Batch least-squares OD is performed using Orekit 13.1 over a 1-day arc, estimating the initial state (initialised from TLE data) and ballistic coefficient. The measurement model uses GNSS positions with assumed accuracy $\sigma = 100$ m and internal 10σ residual rejection. The solution is recomputed independently for each cumulative filtering stage.

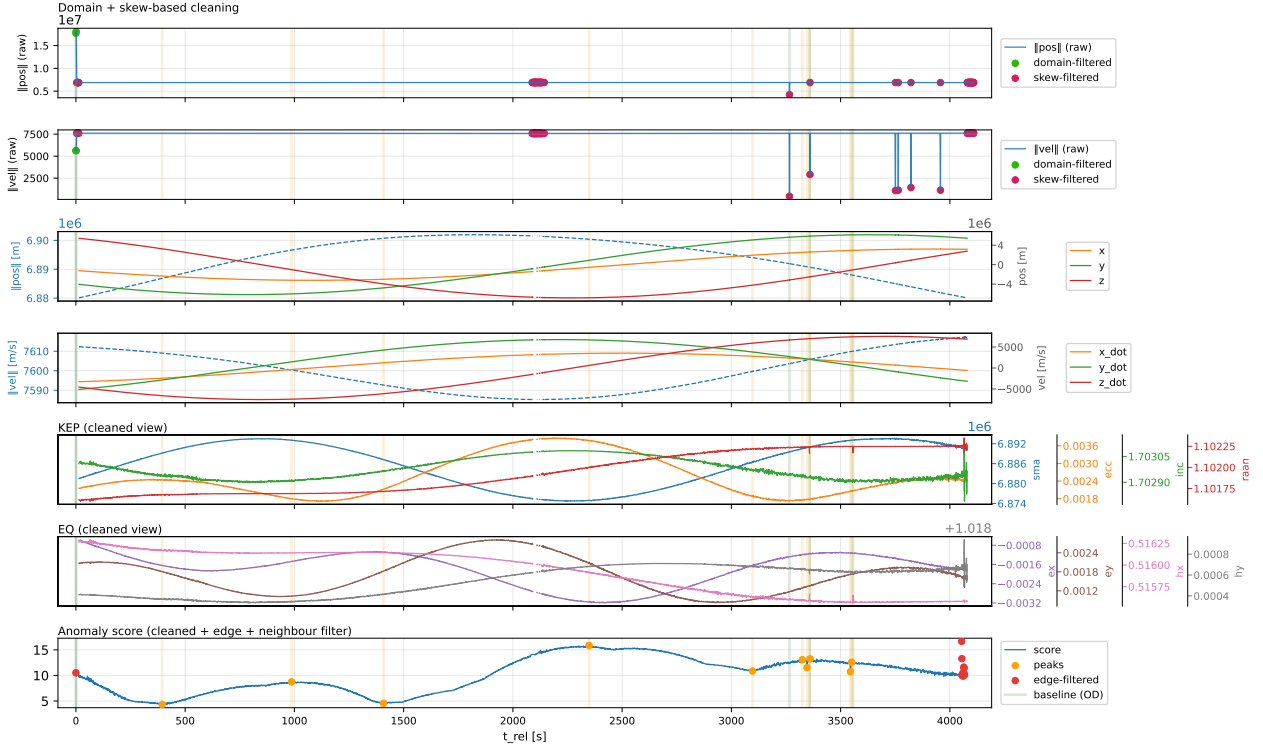


Figure 3: Cleaning applied to a representative segment. Panels 1–2: raw position and velocity magnitudes with samples rejected by domain- and skew-based filters. Panels 3–4: corresponding cleaned magnitudes and components. Panels 5–6: resulting Keplerian and equinoctial elements. Panel 7: anomaly score with detections from the peak, edge, and neighbour stages. Yellow shading: detected peaks. Green shading: measurements rejected by a baseline batch least-squares OD. Both methods agree on obvious outliers, but the proposed pipeline additionally identifies dynamically inconsistent samples that remain undetected by position-only residual analysis.

Stage	WRMS	Δ WRMS	Iter.
original	0.04262	—	5
+ domain & skew	0.01626	−61.85%	5
+ edges	0.01587	−62.76%	4
+ peaks & neighb.	0.01586	−62.79%	4

Table 2: OD performance across cumulative filtering stages. WRMS denotes the weighted root-mean-square of post-fit residuals, computed using the assumed measurement uncertainties as weights and evaluated over the validation interval shown in Figure 3.

the domain- and skew-based stage reduces the **weighted root-mean-square (WRMS)** by approximately 62% relative to the unfiltered case. The subsequent edge and peak stages further stabilise the solution, reaching an overall reduction of about 63%. The magnitude of improvement is scenario-dependent and varies across segments; however, the staged ordering is fixed and reflects progressively stronger assumptions, from physically motivated thresholds to local feature-based corrections.

The observed improvement implies that the proposed filters remove measurements that the baseline OD retains, but which nonetheless degrade the fit quality. In addition,

the number of solver iterations decreases from five to four, indicating improved numerical conditioning.

Across segments, different stages may contribute differently, but the cumulative design ensures conservative behaviour in simple regimes and targeted refinement when structural irregularities are present. The cleaning pipeline therefore acts as an effective preconditioning step, improving solution stability and residual consistency without aggressive data pruning.

5 Conclusion

This work introduces a data-centric preprocessing pipeline for GNSS-based OD that adaptively conditions measurements according to their dynamical structure. By integrating orbital-dynamics-based sanity checks, ML-driven structure assessment, and signal-processing-based anomaly suppression, the proposed approach enables controlled, assumption-aware data refinement without modifying the underlying estimation model.

Across representative operational scenarios, the method consistently reduces residual dispersion, improves numerical conditioning, and shortens convergence while retaining virtually all valid measurements.

The results support the view that robust OD is not solely a modelling problem, but also a data-centric one. Structure-aware preprocessing offers a practical and operationally compatible pathway to improving stability and consistency in real-world GNSS-based OD workflows.

6 Acknowledgments

This work proposes an innovative solution developed within the Competitiveness framework (Element 2) of the European Space Agency's (ESA) Navigation Innovation and Support Program (NAVISP) and was carried out under the Project "Artificial Intelligence Fights Space Debris" N° C626449889-0046305 co-funded by the Recovery and Resilience Plan and NextGeneration EU Funds (www.recuperarportugal.gov.pt). This work is partially supported by NOVA LINC'S (UID/04516/2025 <https://doi.org/10.54499/UID/04516/2025>) with the financial support of FCT/IP.

References

1. Chu, X., Ilyas, I. F., Krishnan, S. & Wang, J. *Data Cleaning: Overview and Emerging Challenges* in *Proceedings of the 2016 International Conference on Management of Data* (Association for Computing Machinery, San Francisco, California, USA, 2016), 2201–2206. ISBN: 9781450335317.
2. Li, P. *et al.* *CleanML: A Study for Evaluating the Impact of Data Cleaning on ML Classification Tasks* in *2021 IEEE 37th International Conference on Data Engineering (ICDE)* (2021), 13–24.
3. Mazumder, M. *et al.* *Dataperf: Benchmarks for Data-Centric AI Development*. *Advances in Neural Information Processing Systems* **36**, 5320–5347 (2023).
4. Siemuri, A., Selvan, K., Kuusniemi, H., Valisuo, P. & Elmusrati, M. S. A Systematic Review of Machine Learning Techniques for GNSS Use Cases. *IEEE Transactions on Aerospace and Electronic Systems* **58**, 5043–5077 (2022).
5. Mohanty, A. & Gao, G. A Survey of Machine Learning Techniques for Improving Global Navigation Satellite Systems. *EURASIP Journal on Advances in Signal Processing* **2024**, 73. ISSN: 1687-6180 (2024).
6. Caldas, F. & Soares, C. Machine Learning in Orbit Estimation: A Survey. *Acta Astronautica* **220**, 97–107. ISSN: 0094-5765 (2024).
7. Bishop, C. M. & Bishop, H. *Deep Learning: Foundations and Concepts* ISBN: 978-3-031-45467-7 (Springer, 2024).
8. Fernández, A. *et al.* *Learning from Imbalanced Data Sets* **2018**. ISBN: 978-3-319-98073-7 (Springer, 2018).
9. Schutz, B., Tapley, B. & Born, G. H. *Statistical Orbit Determination* (Elsevier, 2004).