Assessing the Accuracy of Multitemporal UAV-Derived 3D Models for Cultural Heritage Monitoring

Master Thesis

for the attainment of the Master's degree "Master of Science", abbreviated "MSc"

submitted within the University Master Program for Further Education
"Geographical Information Science & Systems - (UNIGIS MSc)"
at the Department of Geoinformatics - Z_GIS,
Faculty of Digital and Analytical Sciences,
University of Salzburg
submitted by

Luciano Giliberto

Supervisor:
Dr. Christian Neuwirth

Basel, August 2026

Declaration

"I hereby declare that I have written this thesis on my own and that I have not used any sources or other tools and resources other than those indicated. All passages that have been taken literally or in substance from the cited sources are clearly identified as such.

I confirm that I have not used any illustrations from textbooks or other sources unless they are subject to a license permitting free use.

I further confirm that I have used artificial intelligence (AI)-based applications exclusively for those purposes explicitly permitted in the guidelines on the use of generative AI in the UNIGIS program, in its current version. Any deviations from this policy must be explicitly approved in writing by the supervisor. I bear full responsibility for all content of this thesis.
Specifically, AI-based tools (Claude for coding assistance and DeepL for translation and language revision) were used; all AI-generated outputs were manually reviewed and modified to ensure technical correctness.

This thesis has not previously been submitted in the same or a similar form as a Bachelor's/Master's/Diploma thesis/Dissertation/Master Thesis."

Luciano Giliberto

Abstract

The construction of large dams and the artificial reservoirs they create are among the main threats to the archaeological heritage of Western Asia, where fluctuations in water level cyclically expose the sites to erosion. Multitemporal monitoring by UAV photogrammetry can quantify these changes, but its reliability depends critically on georeferencing accuracy, which is often compromised in remote contexts where differential GNSS cannot be used. This thesis evaluates to what extent a low-cost georeferencing workflow - consumer-grade GPS coordinates corrected with a limited number of Total Station points - can produce three-dimensional models with geometric accuracy sufficient for the analysis of multitemporal surface variations. Carried out within the ReLand project at the Gombass site (Mosul Dam reservoir), the study compares six georeferencing scenarios, three coordinate correction models, a Monte Carlo simulation for the minimum number of control points, and a surface comparison based on the M3C2 and DEM of Difference (DoD) algorithms. The pure three-parameter translation proved the most accurate model (mean 3D error of 0.862 m on the independent check points), outperforming the Helmert transformations, which are penalized by the non-separability of the parameters over a small area; seven control points already recover 96% of the planimetric correction. The residual vertical error is dominated by a temporal drift of the elevation bias, which cannot be reduced by adding control points. The surface comparison reveals a systematic tilt of 1.478°, which alone explains 99.5% of the discrepancy; once this component is removed, the random residual is about 0.24 m (NMAD), yielding a minimum level of detection (minLoD) on the order of 0.5 m. The low-cost workflow can therefore reliably detect surface variations of about half a metre or more, while the detection capability in multitemporal monitoring depends on the random residual of the method rather than on the number of control points, because systematic tilt and bias cancel between epochs surveyed with the same method.

Keywords: UAV photogrammetry, low-cost georeferencing, consumer GPS, Total Station, M3C2, DEM of Difference, multitemporal monitoring, archaeological heritage.

Content

List of abbreviations

Abbreviation	Definition
CP	Check Point (independent)
CSV	Comma-Separated Values
DEM	Digital Elevation Model
DGPS	Differential GPS
DoD	DEM of Difference
DSM	Digital Surface Model
EPSG	EPSG Geodetic Parameter Dataset (coordinate reference system code)
EXIF	Exchangeable Image File Format
GCP	Ground Control Point
GNSS	Global Navigation Satellite System
GPS	Global Positioning System
GSD	Ground Sampling Distance
ICP	Iterative Closest Point
IRLS	Iteratively Reweighted Least Squares
LAZ	compressed LAS point cloud format
M3C2	Multiscale Model to Model Cloud Comparison
MAD	Median Absolute Deviation
minLoD	minimum Level of Detection
NMAD	Normalized Median Absolute Deviation
PPK	Post-Processed Kinematic
RMSE	Root Mean Square Error
RTK	Real-Time Kinematic
SD	Standard Deviation
SfM	Structure from Motion
SVD	Singular Value Decomposition
TS	Total Station
UAV	Unmanned Aerial Vehicle

Abbreviation	Definition
UTC	Coordinated Universal Time
UTM	Universal Transverse Mercator
WGS84	World Geodetic System 1984
Tell	(archaeology) artificial mound formed by the stratified accumulation of successive settlements; here it denotes the archaeological site of Gombass

List of Figures

List of Tables

List of Formulas

1. Introduction

The construction of large dams and the formation of artificial reservoirs represent one of the most widespread threats to the cultural heritage of Western Asia. A single reservoir may permanently or periodically submerge hundreds of archaeological sites, many of which are only partially documented before inundation (Cunliffe et al., 2012; Marchetti et al., 2019; Sconzo et al., 2023; Sconzo & Simi, 2024). Water-level fluctuations, driven by reservoir management and changing environmental conditions, repeatedly expose and submerge archaeological deposits, accelerating erosion, sediment reworking, and the progressive loss of archaeological material and contextual information (Marchetti et al., 2019; Sconzo et al., 2023, with further literature). Consequently, these rapidly changing landscapes require repeated documentation and long-term monitoring in order to assess their evolution and support conservation strategies.

Unmanned aerial vehicle (UAV) photogrammetry based on Structure-from-Motion (SfM) algorithms has become a key tool for the high-resolution documentation of archaeological sites and landscapes. Compared with traditional surveying techniques, it enables the rapid and cost-effective production of orthophotos, digital surface models (DSMs), and detailed three-dimensional reconstructions (Campana, 2017; Fiz et al., 2022; Puniach et al., 2026). Repeated UAV surveys can be compared through point-cloud analysis and DEM of Difference (DoD) techniques, making it possible to quantify surface changes and estimate volumetric loss over time (Lague et al., 2013; Wheaton et al., 2010; He et al., 2025). However, these analyses are meaningful only if the models acquired during different campaigns are geometrically comparable.

The reliability of multitemporal comparison therefore depends critically on georeferencing accuracy. Systematic positional differences between surveys may generate apparent surface changes that are indistinguishable from those produced by real erosion (Liu et al., 2022; Zhong et al., 2025).

This issue is particularly relevant in remote archaeological contexts, where logistical constraints and regulatory limitations often prevent the routine use of differential GNSS and other high-precision geodetic techniques (Herrmann et al., 2018). Under these conditions, Ground Control Points (GCPs) are commonly surveyed using consumer-grade GPS receivers, whose coordinates are affected by systematic planimetric and vertical errors, reducing the absolute geometric accuracy of the resulting photogrammetric models (Kim et al., 2023; Štroner et al., 2021).

A practical alternative consists of combining consumer-grade GPS measurements with a limited number of high-accuracy reference points surveyed with a Total Station. Such an approach has the potential to improve the geometric consistency of multitemporal datasets while remaining compatible with the logistical constraints of archaeological fieldwork. Nevertheless, the effectiveness of different correction models, the minimum number of Total Station reference points required for a stable solution, and the geometric accuracy achievable for multitemporal surface-change analysis have not yet been systematically evaluated.

This thesis was developed within the framework of the ReLand (Resurfacing Landscapes of the Mosul Dam Reservoir) project, an international research initiative dedicated to the monitoring and documentation of the archaeological heritage threatened by the Mosul Dam reservoir in northern Iraq (Sconzo, 2024; Sconzo et al., 2023).

The research investigates the following question:

To what extent can a low-cost georeferencing workflow, combining consumer-grade GPS measurements with a limited number of Total Station reference points, produce UAV-derived three-dimensional models with sufficient geometric accuracy for the analysis of multitemporal surface variations under real archaeological field conditions?

To answer this question, the study pursues four objectives:

- to characterize the positional error of consumer-grade GPS coordinates with respect to the Total Station reference;

- to compare coordinate correction models of increasing complexity and identify the most accurate solution;
- to determine, through Monte Carlo simulation, the minimum number of Total Station reference points required for a stable correction;
- to evaluate the geometric accuracy of the resulting photogrammetric models and quantify the minimum level of detection for reliable multitemporal surface-change analysis.

The proposed workflow is evaluated through six georeferencing scenarios derived from a UAV survey acquired in November 2025. The scenarios include the Total Station reference (S1), the gold standard against which all the others are evaluated; the original consumer GPS solution (S2); three coordinate correction models of increasing complexity (S3-S5); and a reduced-control-point configuration (S6) defined through Monte Carlo simulation. Their performance is assessed both at the coordinate level, using independent check points, and at the surface level through M3C2 and DEM of Difference analyses.

The remainder of this thesis is organized as follows. Chapter 2 describes the study area, datasets, and methodological workflow, including the GPS error analysis, coordinate correction models, Monte Carlo simulation, photogrammetric processing, and surface comparison. Chapter 3 presents the results. Chapter 4 discusses the findings in relation to the research question and previous literature. Finally, Chapter 5 summarizes the main conclusions, methodological limitations, and directions for future research.

2. Materials and Methods

2.1 Study area and available datasets

This section introduces the study area and the datasets used throughout the thesis.

Section 2.1.1 describes the archaeological and geomorphological context of the Mosul Dam reservoir, while Section 2.1.2 presents the rationale for selecting the study area and the datasets employed: the UAV image block georeferenced with consumer-grade

GPS and the Ground Control Points (GCPs) measured with both consumer-grade GPS and a Total Station.

The methodological workflow adopted in this study is summarized in Figure 2.1 and consists of six sequential phases:

- Data acquisition, including the UAV image block and the twenty-three GCPs surveyed with both consumer-grade GPS and Total Station (§2.1.2);
- GPS error analysis and coordinate correction, where the positional error is characterised and different correction models are estimated (§2.3);
- Monte Carlo simulation, used to determine the minimum number of Total Station reference points required for a stable correction (§2.4);
- Photogrammetric reconstruction, in which the different georeferencing scenarios are processed in Agisoft Metashape to generate dense point clouds and digital elevation models (§2.5.1);
- Surface comparison, performed in CloudCompare using M3C2 and DEM of Difference (DoD) analyses after rigid ICP registration (§2.5.3-2.5.5);
- Accuracy assessment, where NMAD and the minimum Level of Detection (minLoD) are computed to evaluate the geometric reliability of the workflow for multitemporal surface-change analysis (§2.5.5).

The output of each phase provides the input for the next one: corrected coordinates are used for photogrammetric reconstruction, the reconstructed surfaces are compared to quantify geometric discrepancies, and the resulting residuals are analysed to estimate the minimum detectable threshold for future multitemporal change.

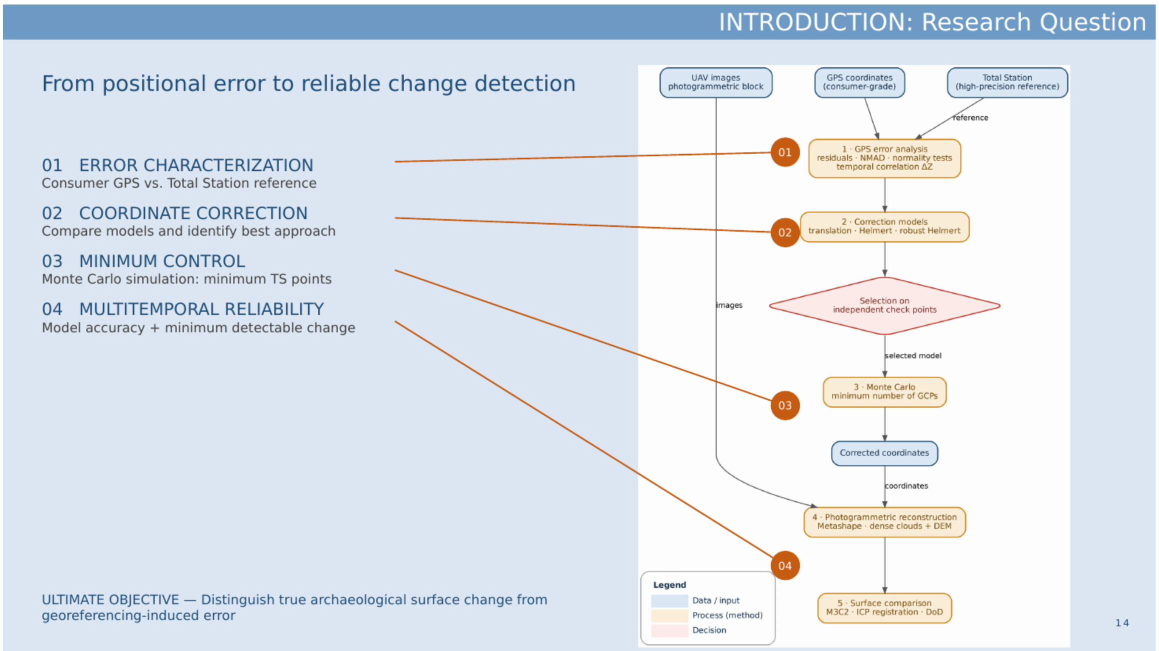

Figure 2.1. Overview of the methodological workflow.

2.1.1 Archaeological and geomorphological setting

The investigated site belongs to the archaeological landscape of the Mosul Dam reservoir, one of the most representative contexts in Western Asia for investigating the impact of large hydraulic infrastructures on cultural heritage. Built on the Upper Tigris between 1981 and 1986, the dam created a reservoir of approximately 372 km² - the largest in Iraq and the fourth largest in the Middle East - which permanently submerged a long stretch of the river valley and around 280 archaeological sites (Sconzo & Simi, 2020; Sconzo et al., 2023; Sconzo, 2024) (Figure 2.2).

Before inundation, the rescue survey conducted in 1980 documented 149 archaeological sites distributed across the basin, while the excavations carried out during the construction of the dam revealed occupation sequences ranging from the Neolithic to the Islamic period (Sconzo & Simi, 2020). More recent regional surveys have further demonstrated the exceptional archaeological richness of the Upper Tigris, highlighting a settlement density far greater than previously assumed (Sconzo, 2024; Sconzo et al., 2023).

Figure 2.2. Location of the study area on the Mosul Dam reservoir (Upper Tigris, Iraq; 36°43’39”N, 42°53’52”E, UTM zone 38N). Author's elaboration on Google Satellite imagery (© Google), accessed 08/08/2026. The red circle marks the study area; the inset shows the regional setting.

Unlike permanently inundated reservoirs, the Mosul basin is characterized by pronounced water-level fluctuations driven by reservoir management and climatic variability. Since the first impoundment, the lake level has varied by more than 40 m, periodically exposing archaeological sites located within the drawdown zone. Each re-emergence provides an opportunity for renewed archaeological investigation, but also exposes the sites to rapid erosion, sediment reworking, and looting. Consequently, repeated documentation has become a fundamental requirement for both archaeological research and heritage management.

These activities are carried out within the framework of the ReLand (Resurfacing Landscapes of the Mosul Dam Reservoir) project, launched in 2023 through the collaboration between the University of Palermo and the Directorate of Antiquities of Duhok. The project aims to assess the effects of reservoir fluctuations on the archaeological landscape of the Upper Tigris and to develop practical protocols for documentation, monitoring, and conservation of threatened cultural heritage (Sconzo, 2024).

From a geomorphological perspective, the study area lies on the north-western margin of the Zagros Mountains and is characterized by sedimentary formations, semi-arid climatic conditions, and active erosional processes (Forti et al., 2021, 2022, 2023a). The investigated sites are tells-anthropogenic mounds formed by the long-term accumulation of successive human occupations (Rosen, 1986; Wilkinson, 2003). Their steep slopes and unconsolidated archaeological deposits make them particularly vulnerable to water erosion, favouring the development of rills and gullies during repeated exposure to reservoir fluctuations (Forti et al., 2023b, 2023c). As a consequence, changes in surface morphology and volume provide an effective proxy for evaluating the progressive loss of archaeological material.

2.1.2 Site selection criteria and available datasets

The study area was deliberately selected to represent a worst-case scenario for evaluating the proposed low-cost georeferencing workflow. Rather than testing the methodology under favourable conditions, the objective was to assess its performance on a large and geomorphologically complex archaeological landscape, representative of the operational challenges commonly encountered during fieldwork in the Mosul Dam reservoir. The investigated area comprises two adjacent tells - Gombass (ReL_16) and ReL_47 - separated by a deeply incised wadi and treated as a single study area (Figure 2.3).

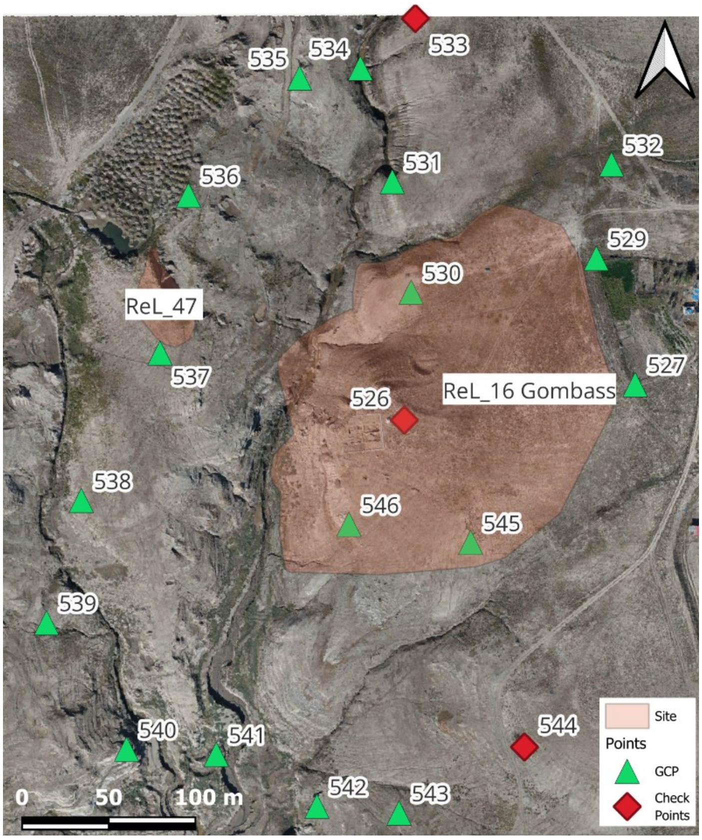

Figure 2.3. Study area showing the two tells Gombass (ReL_16) and ReL_47 with the surveyed points (©ReLand)

Together, they cover approximately 470 × 420 m (19.74 ha), making this area larger than any individual archaeological site documented within the ReLand survey (Table A10). By comparison, most archaeological sites in the reservoir are considerably smaller: 37% occupy less than 0.5 ha, whereas only 10% exceed 10 ha (Table 2.1).

Figure 2.4. View of tell Gombass, from North(©ReLand)

Class (ha)	N sites	% sites	Area (ha)	% area
< 0.5	32	37.2	6.40	2.4
0.5-1	6	7.0	5.00	1.9
1-2	15	17.4	20.47	7.8
2-5	19	22.1	64.65	24.7
5-10	5	5.8	32.39	12.4
> 10	9	10.5	132.45	50.7
Total	86	100	261.36	100

Table 2.1. Distribution of sites by size class, with the percentage share of the number of sites and of the total area. The classification covers the 86 sites surveyed by the ReLand mission. The 2-5 ha class includes the Gombass site (3.13 ha).

Beyond its size, the study area presents several characteristics that increase the complexity of both field surveying and photogrammetric reconstruction. The two mounds are separated by an irregular wadi, exhibit marked elevation differences and show widespread erosion features produced by repeated exposure to reservoir fluctuations. These factors complicate the distribution of Ground Control Points (GCPs), increase the extent of the UAV survey, and provide a demanding test for the geometric consistency of the proposed workflow.

The same complexity was reflected during field operations. Establishing and measuring control points over such a large and articulated area required long acquisition times using consumer-grade GPS equipment, reproducing the practical constraints typical of archaeological surveys conducted under limited logistical resources.

The acquisition strategy directly reflects the methodological objective of this thesis. Since the aim is to evaluate whether a low-cost georeferencing workflow can produce UAV-derived models suitable for multitemporal surface-change analysis, the study relies on two complementary datasets:

- a UAV image block (869 photos), whose Ground Control Points are measured with a consumer-grade GPS receiver (Garmin GPSMAP 65s), reproducing the operational conditions of a low-cost archaeological survey;
- the same points re-measured with a Total Station (±5 mm), providing the high-accuracy geometric reference used to estimate and validate the correction of the consumer-grade GPS coordinates.

The UAV survey was carried out on 20-21 November 2025 within the ReLand project using two DJI Air 3S platforms, each assigned to one of two orthogonal flight grids (East-West and North-South). Owing to the extent of the site, each grid required three consecutive flights, resulting in a total of six acquisition sessions. The present study uses only the East-West image block (869 images), acquired at an altitude of 65 m, because it provides a finer Ground Sampling Distance (GSD) and more homogeneous illumination than the North-South block (1,055 images acquired at 80 m). The latter dataset is reserved for future multitemporal investigations.

Ground control measurements were acquired during the same field campaign. Twenty-three Ground Control Points were surveyed both with a Garmin GPSMAP 65s receiver and with a Trimble C5 Total Station. Measuring the same points with both instruments provides the basis for characterising the positional error of the consumer-grade GPS observations and for estimating the correction models described in the following sections. GPS observations were collected using repeated position averaging (30-60 fixes per point), while the receiver was mounted on a levelling pole to ensure a constant antenna height and vertical alignment throughout the survey.

Figure 2.5. Control-point measurement with the consumer-grade GPS receiver on a levelling pole (©ReLand).

The Total Station survey was performed on the same day using a free-station configuration referenced to three differential GNSS control points. With a nominal positional accuracy of ±5 mm, the Total Station dataset is adopted throughout this study as the geometric reference against which all consumer-grade GPS solutions are evaluated.

All coordinates are expressed in the WGS 84 / UTM Zone 38N reference system (EPSG:32638). The common set of twenty-three surveyed points constitutes the basis for estimating and validating the GPS correction models, whereas the East-West UAV image block provides the input dataset for the photogrammetric reconstruction. The functional classification of the control points and the definition of the six georeferencing scenarios are described in the following sections.

2.2 Ground control measurements and georeferencing scenarios

Once the datasets have been defined, this section establishes the experimental setup of the whole evaluation. It describes how the twenty-three common points are assigned to three categories - control points, independent check points, and excluded points - and defines the six georeferencing scenarios, which are distinguished by the source and the processing of the GCP coordinates (Table 2.2). These scenarios are the common reference of the error analysis and the correction models (§ 2.3), of the Monte Carlo analysis (§ 2.4), and of the photogrammetric processing (§ 2.5).

The 23 common points were divided into three categories. Nineteen form the pool of Ground Control Points (GCP) used to estimate the correction parameters. Three points (ID 526, 533, 544), in the central-western, northern, and south-eastern sectors of the area, serve as independent check points (CP). They never enter the parameter estimation, so as to guarantee an unbiased evaluation of the residual accuracy, following the standard procedure of Höhle and Höhle (2009, p. 399). One point (ID 529) is excluded from all analyses, because it is poorly visible in the images and its manual projection in the photographs is therefore unreliable.

The GPS measurements used multiple averaging, with 30-60 fixes per point and the mean automatically computed by the receiver firmware. The session lasted about 1.5 hours (06:16-07:53 UTC), with the points measured in spatial sequence along the study area. The Total Station measurements were carried out on the same day, with a free-station survey oriented on three DGPS reference points. This brings the coordinates into the UTM 38N system.

The six georeferencing scenarios for processing in Agisoft Metashape are distinguished only by the source and the processing of the GCP coordinates (Table 2.2). Scenario S1 uses the Total Station coordinates directly and is the geometric reference against which all the others are evaluated. Scenarios S2-S5 use the raw consumer GPS coordinates (S2) or coordinates corrected with one of the three transformation models described in § 2.3 (S3-S5). In all scenarios the three check points retain the Total Station coordinates with an accuracy of 0.005 m. The GCPs of the GPS scenarios are instead assigned an asymmetric accuracy - 1.000 m in planimetry (XY) and 20.000 m in elevation (Z). This choice reflects the residual error expected after the correction and avoids over-constraining the absolute elevation in the bundle adjustment; the rationale is in § 2.5.1. Scenario S6 is identical in method to the best among scenarios S3-S5, but is estimated on a reduced GCP configuration, whose size is defined by the Monte Carlo analysis (§ 3.3). It serves to validate at the model level the reduction of the number of control points.

Scenario	GCP coordinate source	n GCP	n CP	Accuracy XY / Z (m)
S1 - TS (reference)	Total Station (±5 mm)	19	3	0.005 / 0.005
S2 - Raw GPS	Consumer GPS (no correction)	19	3	1.000 / 20.000
S3 - Translation	GPS + 3-parameter translation	19	3	1.000 / 20.000
S4 - Helmert	GPS + 7-parameter Helmert	19	3	1.000 / 20.000
S5 - Robust Helmert	GPS + Helmert (Huber k=1.345)	19	3	1.000 / 20.000
S6 - Translation, reduced GCP	GPS + (Best between S3-S5), reduced GCP subset	reduced (§ 3.3)	3	1.000 / 20.000

Table 2.2. Georeferencing scenarios evaluated in this study. GCP = Ground Control Points; CP = independent check points (ID 526, 533, 544). Point ID 529 is excluded from all scenarios. The reported accuracy refers to the value assigned in Agisoft Metashape for each point category, separately for the planimetric (XY) and vertical (Z) components. (Note: The deliberately loose Z accuracy in the GPS scenarios prevents over-constraining the absolute elevation in the bundle adjustment (cf. § 2.5.1). For the TS check points the value is 0.005 m in all scenarios).

2.3 GPS error analysis and coordinate correction models

This section characterizes the error of the consumer-grade GPS coordinates with respect to the Total Station reference and defines the models used to correct it. § 2.3.1 describes its statistical structure with an analysis of the residuals by component; § 2.3.2 motivates, on this basis, the choice of the most appropriate transformation model; § 2.3.3 presents the three correction models actually implemented, which generate the input coordinates for the corrected GPS scenarios (S3-S5). This characterization also provides the interpretive basis for the results of § 3.1 and § 3.2.

2.3.1 Descriptive analysis of the positional residuals

To define an effective correction model, one must first understand the nature of the error to be corrected. A positional residual can reflect very different structures - a constant systematic bias, Gaussian random noise, a time- or space-dependent drift, or anomalous values - and each requires a different statistical treatment and correction model. For this reason the descriptive analysis of the residuals does not stop at the conventional accuracy metrics, but complements them with diagnostic tools that reveal their underlying structure. In particular, robust estimators are used alongside the classical ones. This responds to the need, emphasized by Höhle and Höhle (2009) for the accuracy assessment of digital elevation models, not to let a few extreme values distort the estimate of the typical dispersion of the error. The metrics and tests described below are therefore chosen for this double purpose, descriptive and diagnostic.

For each of the 23 common points, the residuals between the consumer GPS position and the Total Station reference position are computed, component by component, according to the relations:

$$\Delta E = E_{\mathrm{GPS}} - E_{\mathrm{TS}}, \quad \Delta N = N_{\mathrm{GPS}} - N_{\mathrm{TS}}, \quad \Delta Z = Z_{\mathrm{GPS}} - Z_{\mathrm{TS}}$$

Formula 2.1. Per-component GPS-TS residuals

The planimetric distance and the three-dimensional distance are computed for each point:

$$d_{2D} = \sqrt{\Delta E^2 + \Delta N^2}$$

Formula 2.2. Planimetric distance

$$d_{3D} = \sqrt{\Delta E^2 + \Delta N^2 + \Delta Z^2}$$

Formula 2.3. Three-dimensional distance

For each component and distance, the descriptive statistics are computed: mean, median, standard deviation (SD), minimum, maximum, and Root Mean Square Error (RMSE), the standard metric for accuracy assessment in photogrammetry and geodesy (Höhle & Höhle, 2009). The same authors (2009, p. 400) also recommend the Normalized Median Absolute Deviation (NMAD) as a robust measure of dispersion, an alternative to SD because it is insensitive to outliers:

$$\mathrm{NMAD} = 1.4826 \cdot \mathrm{median}|\, x_i - \mathrm{median}(x)\,|$$

Formula 2.4. NMAD (Normalized Median Absolute Deviation)

An SD/NMAD ratio > 1 signals extreme values that inflate the SD with respect to the typical dispersion. The computed values of the descriptive and robust statistics are reported in Tables 3.1 and 3.2 (§ 3.1).

The compatibility of the residuals with the Gaussian hypothesis is verified with the Shapiro-Wilk test (Shapiro & Wilk, 1965) (Table 3.5). The temporal structure of the elevation error is investigated with a Pearson correlation test (Table 3.6) between UTC acquisition time and ΔZ. The spatial structure of the error is explored with omnidirectional empirical semivariograms (Gräler et al., 2016) and with the analysis of the correlation between the planimetric residual d2D and the distance from the nearest GCP, in line with Villanueva and Blanco (2019, p. 172). All the analyses are implemented in R (R Core Team, 2024) following the principles of Wilson et al. (2017).

Three diagnostic tools are used, and each deserves a brief explanation. The first is the NMAD (normalized median absolute deviation), a measure of dispersion built to resist anomalous values. Instead of starting from the mean, as the standard deviation does, it starts from the median and looks at how much the individual residuals depart from it in absolute value. It takes the median of these deviations and multiplies it by 1.4826, so that on well-behaved data the result coincides with the standard deviation. Because it is based on medians, a few extreme points hardly alter it at all (Höhle & Höhle, 2009). The second is the Shapiro-Wilk test, which answers the question “do these residuals resemble a normal, bell-shaped distribution?”. A low probability indicates that they do not, and therefore that the classical mean and standard deviation can mislead and must be complemented by robust measures (Shapiro & Wilk, 1965). The third is the semivariogram, a simple way to check whether nearby points tend to have similar errors: one observes how the typical difference between two measurements grows with their distance. A flat trend indicates that the error has no spatial structure - it is in practice random in space - while an increasing trend indicates that proximity matters (Gräler et al., 2016). These tools are chosen precisely because the GPS error may not be Gaussian and may contain a systematic part: measures are needed that are not fooled by outliers and that reveal any hidden structure.

2.3.2 Selection of the transformation model

The transformation models between coordinate systems have increasing complexity: from the rigid translation, to the seven-parameter similarity transformation, up to non-rigid models with more degrees of freedom. The choice is not without consequence. A more complex model always reduces the residual on the estimation points, but only if the geometry of the control network allows its additional parameters to be estimated independently. Otherwise these capture the random noise instead of the real geometry of the terrain, and the model predicts worse on the independent points (Li et al., 2012; Tran et al., 2023). The model selection was therefore guided, even before the performance comparison, by a preliminary analysis of the separability of the parameters on the specific geometry of the site.

The choice of the most appropriate model was based on a preliminary analysis of the variability of the scale factor between the two measurement systems (Table 3.4). For all the available point pairs, the ratio between the inter-point distances in the GPS system and in the TS system was computed. The distribution of these ratios is reported in § 3.2. The question is one of identifiability. If such variability reflects the noise of the consumer GPS rather than a real scale difference between the two systems, the rotation and scale parameters of the seven-parameter Helmert transformation are not reliably separable.

Over an area of about 500 × 500 m, a rotation of 0.001° applied to UTM coordinates of the order of 4 × 10^6 m produces apparent translations of tens of metres, indistinguishable from the real translational component. Li et al. (2012, p. 2100) document the same instability of the traditional seven-parameter model when the common points do not allow the transformation components to be separated independently. For this reason the most parsimonious correction model is adopted as the reference, while the Helmert models are evaluated as methodological comparisons.

2.3.3 Implemented correction models

Implementing a graduated family of models, instead of a single estimator, serves two purposes. On the one hand it allows a controlled comparison, in which each model adds a single element of complexity with respect to the previous one, isolating its contribution to accuracy. On the other hand it covers the different hypotheses on the structure of the GPS error. These range from pure translation, which assumes a constant bias over the whole field, to robust estimators, which admit anomalous observations of heterogeneous quality (Höhle & Höhle, 2009; Tran et al., 2023). All the closed-form estimates are based on rigid alignment via singular value decomposition (Eggert et al., 1997), which centres the coordinates on their respective centroids to guarantee numerical stability. The framework of the weighted symmetric Helmert transformation follows Mercan et al. (2018). The three models are described below in order of increasing complexity.

Three transformation models of increasing complexity were implemented in R and applied to the consumer GPS coordinates, to generate the input files of the three corrected GPS scenarios (S3-S5).

Scenario S3 - Pure three-parameter translation. The translation is the difference between the centroid of the TS coordinates and that of the GPS coordinates, computed on the 19 available GCPs:

$$\mathbf{t} = \bar{\mathbf{X}}_{\mathrm{TS}} - \bar{\mathbf{X}}_{\mathrm{GPS}}$$

Formula 2.5. Three-parameter translation vector where X is the vector of the coordinates (E, N, Z). It is the least-squares solution of the pure translation model: it minimizes the sum of the squares of the residuals on the control points (Eggert et al., 1997, p. 273). The same translation is then applied to all the GPS points:

$$\mathbf{X}'_{\mathrm{GPS}} = \mathbf{X}_{\mathrm{GPS}} + \mathbf{t}$$

Formula 2.6. Translation applied to the GPS coordinates

The estimated parameters are reported in § 3.2.

Scenario S4 - Seven-parameter Helmert transformation with local coordinates. The Helmert model estimates three translations, three rotations, and a scale factor. The absolute UTM coordinates ill-condition the covariance matrix and make the estimate unstable. To avoid this, the estimation uses local coordinates, obtained by subtracting from all the coordinates the centroid of the TS dataset, following the rigid alignment procedure of Eggert et al. (1997, pp. 273-274). The rotation matrix and the scale factor derive from the corresponding closed-form solution based on singular value decomposition (SVD). The framework of the weighted symmetric Helmert transformation follows Mercan et al. (2018). The estimated parameters are reported in § 3.2.

Scenario S5 - Helmert with Huber robust estimator. S5 extends S4 with an iteratively reweighted least squares (IRLS) estimation based on the global Huber M-estimator, which automatically identifies and reduces the contribution of the less accurate GPS points. The Huber weight function assigns unit weight to points with normalized residual $|u| \leq k$ (tuning constant $k = 1.345$) and weight $k/|u|$ to points with $|u| > k$. Here σ is estimated as median absolute deviation (MAD)/0.6745 (Höhle & Höhle, 2009, p. 400):

$$w(u) = \begin{cases} 1, & u \leq 1 \\ 1/u, & u > 1 \end{cases} \qquad u = \frac{r}{k\,\sigma}, \quad k = 1.345$$

Formula 2.7. Huber weight function

The estimation proceeds iteratively until convergence.

Each model works in a different way and was included for a specific reason. The translation (S3) is the simplest correction: it shifts all the points by the same fixed amount in X, Y, and Z, equal to the mean residual between GPS and Total Station on the control points. That is, it assumes that the GPS error is essentially a constant offset. Having very few parameters, it is difficult to overfit. The Helmert transformation (S4) adds to the translation three small rotations and a scale factor - seven parameters in all - in the Bursa-Wolf form (Tran et al., 2023, p. 126, Eq. 1). The old coordinates are rotated by three minimal angles, multiplied by a scale factor (1+k), and finally translated; at least three common points are needed to solve it. It can absorb a slight tilt or dilation of the system, but over a small area these additional parameters are difficult to distinguish from the simple translation. The robust estimator (S5 Huber) does not change the model, but the way its parameters are estimated. Instead of weighting every point equally, it reduces the influence of the points that deviate greatly from the others (the outliers). Huber is a classical M-estimator that gradually attenuates the large residuals (Huber, 1992). The robust estimator is effective when a few points are of poor quality. Here, however, the dominant vertical error might be a systematic drift - valid information, not the failure of a few points for which it is designed. Reporting all three models serves precisely to verify whether the robust estimation brings a benefit, without taking it for granted.

2.4 Monte Carlo analysis for the minimum number of GCPs

The Monte Carlo analysis quantifies the minimum number of Total Station points needed for a stable correction and characterizes its statistical variability. The procedure rests on the iterative resampling of the GCP pool and on the independent evaluation on the check points. It is applied to the correction model adopted as the reference in § 2.3 and is its operational complement: it translates that accuracy into a recommendation on the number of control points, whose outcomes are reported in § 3.3.

The Monte Carlo analysis quantifies the minimum number of Total Station reference points needed for a stable and geometrically reliable correction of the GPS coordinates. It also characterizes how the RMSE varies with the size of the estimation sample. The methodology is conceptually analogous to that of Villanueva and Blanco (2019, p. 168), who assess the effect of the number and distribution of the GCPs on the accuracy of the UAV DEM through iterative simulations. Here it is extended to the specific case of GPS coordinate correction by translation.

The Monte Carlo approach is suitable for this problem because the relationship between the number of control points and accuracy has no closed analytical form. It depends on the spatial configuration of the extracted points, on the error structure, and on their interaction - factors that are better evaluated by simulation than theoretically. The literature on GCP configuration in UAV photogrammetry has consolidated precisely this type of analysis by iterative subsampling. Subsets of control points of variable size are repeatedly extracted, and the performance is evaluated on independent points (Sanz-Ablanedo et al., 2018; Villanueva & Blanco, 2019; Zhong et al., 2025). Transferring this scheme from bundle adjustment estimation to the estimation of the GPS coordinate correction parameters preserves its logic. Evaluating on points not used in the estimation gives a measure of accuracy uncontaminated by the fit to the data, while repeating over many realizations with a fixed seed ensures both stability and reproducibility (Wilson et al., 2017).

For each value of n (number of GCPs used for the estimation of the translation, with $n \in [3, 19]$), 500 independent iterations are performed with a fixed seed (set.seed(42)). This makes the results reproducible, following the principles of Wilson et al. (2017). At each iteration, n points are drawn at random, without replacement, from the pool of the 19 available GCPs. The translation is estimated on these n points and applied to the 3 fixed check points (ID 526, 533, 544), and the planimetric (RMSE_2D), elevation (RMSE_Z), and three-dimensional (RMSE_3D) RMSE is computed on them. Measuring accuracy on the check points rather than on the estimation points gives an independent evaluation, uncontaminated by the fit of the model to the training data, as recommended by Sanz-Ablanedo et al. (2018). For each value of n, the summary statistics are then computed: median, 5th and 95th percentile of the RMSE over the 500 iterations. The total realizations are $17 \times 500 = 8{,}500$. The 500 iterations keep the percentiles stable even for the smallest samples ($n = 3$).

The “knee” of the curve - the value of n beyond which the improvement of the median RMSE becomes negligible - was identified by computing the percentage improvement for each additional GCP. The analysis of the 5-95% confidence band also makes it possible to distinguish two different benefits of increasing the GCPs: the reduction of the median RMSE and the reduction of the width of the band, that is, of the risk of poor results due to an unfavourable selection of the sample. This offers a quantitative basis for the operational recommendations.

2.5 Photogrammetric processing and model comparison

This phase moves from the corrected coordinates to the three-dimensional models and their comparison, where the accuracy of the georeferencing is evaluated at the surface level. The alignment and the photogrammetric processing in Agisoft Metashape are treated in § 2.5.1. § 2.5.2 explains why a surface-level comparison is needed in addition to the one on the coordinates. The M3C2 algorithm and the decomposition of the systematic error component are covered in §§ 2.5.3 and 2.5.4. § 2.5.5 defines the DEM of Difference, the minimum level of detection, and the volumetric estimate, and § 2.5.6 specifies the design of the scenario comparisons. The quantitative outcomes of this phase are reported in § 3.4.

2.5.1 Alignment and photogrammetric processing

The photogrammetric processing was carried out with Agisoft Metashape Professional (version 2.x). The image alignment was conducted with High accuracy, tie point limit set to 0 (unlimited), Adaptive Camera Model Fitting active, and Exclude Stationary Tie Points active. The East-West flight group was processed as the main chunk, on account of its greater temporal and radiometric coherence with respect to the North-South group. The quality of the sparse cloud was improved with an iterative filtering (Gradual Selection tool) applied in sequence to three criteria - Reconstruction Uncertainty (threshold 10), Projection Accuracy (threshold 3), and Reprojection Error (threshold 0.3). Each step was followed by a Camera Optimisation with all the intrinsic parameters active (f, cx, cy, k1-k4, p1-p2, b1-b2, Adaptive Camera Model Fitting). This cleaning workflow follows the

USGS guidelines (Over et al., 2021); the resulting mean reprojection error is reported in § 3.2.

Not all the scenarios are carried through to the full photogrammetric reconstruction. Those that are include the Total Station reference scenario, the main correction scenario, and the reduced control point configuration; the others are evaluated only at the level of the corrected coordinates. The selection criteria and the corresponding outcomes are in § 3.2 and § 3.3. To guarantee the full comparability among the three models, the alignment and the cleaning of the sparse cloud (Gradual Selection) were performed only once on a single base chunk. This was then duplicated for each scenario by replacing only the marker coordinates. For each scenario, the CSV file with the corrected GCP coordinates generated by the R scripts was imported into Metashape (Reference → Import). The check points were kept as verification points, not used in the bundle adjustment.

A critical point of the workflow is the treatment of the camera GPS positions in the georeferenced bundle adjustment. The GPS elevations of the DJI Air 3S are systematically too low by about 100 m with respect to the terrain. The block comprises 869 cameras, each with a default elevation accuracy (10 m) more stringent than that of the GPS markers (20 m, cf. § 2.2). For this reason the camera observations would numerically dominate the 19 markers and would drag the model towards the wrong elevation, systematically tilting the reconstructed surface. The camera positions were therefore disabled from the optimization (Reference) in all scenarios, in an identical way, leaving only the markers to constrain the vertical georeferencing. In S1 the Total Station markers (accuracy 0.005 m) prevail over the cameras anyway, but the deactivation is maintained here too for pipeline consistency. The planimetry is not affected by the problem, because in XY the markers (1 m) prevail over the cameras. The accuracy of the GPS markers is deliberately asymmetric - 1.000 m in XY and 20.000 m in Z - so that the absolute elevation remains weakly constrained and avoids the deformation of the model. This comes at the cost of a vertical residual of a few metres, documented as a distinct result in § 3.2.

After importing the markers, a Camera Optimisation was performed with the cameras disabled, optimizing the intrinsic parameters (f, cx, cy, k1-k4, p1, p2, b1, b2, Adaptive Camera Model Fitting active, Fit additional corrections disabled). The accuracy of each scenario was then evaluated with the RMSE on the three independent check points. The dense cloud was generated with Medium quality and Mild depth filtering. The choice of Medium - instead of High - deliberately reflects the computational constraints of a low-cost operational workflow, consistently with the research question, which concerns the accuracy attainable in real operational conditions with limited resources and not in optimal laboratory conditions. The same setting is used for all the scenarios, so as not to introduce processing differences extraneous to the georeferencing. The Digital Elevation Models (DEM) were generated from the dense cloud and exported in GeoTIFF at the native resolution for each processed scenario. The dense clouds were exported in LAZ format (EPSG:32638) for the comparison in CloudCompare.

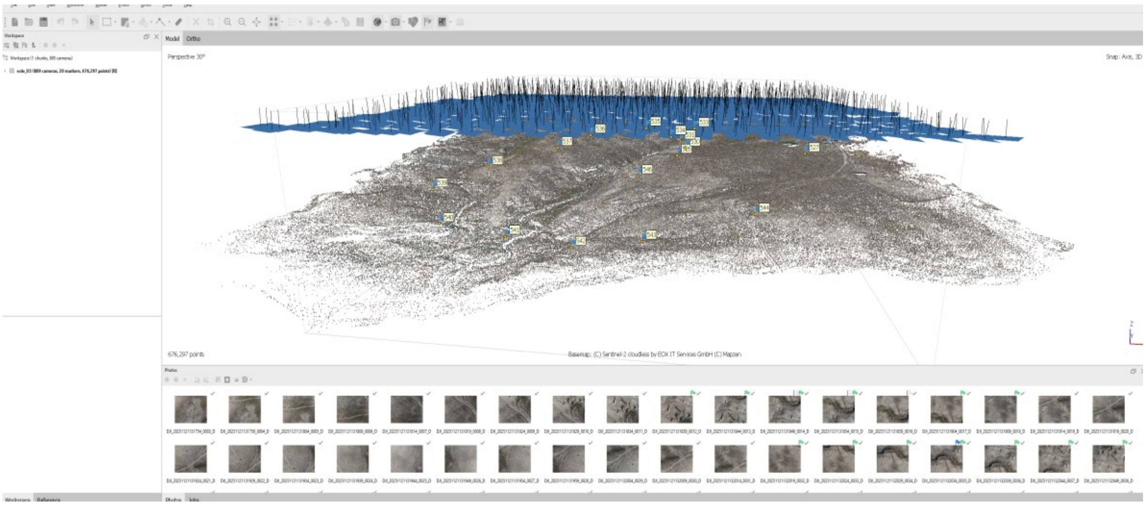

Figure 2.6. Photogrammetric reconstruction of the study area in Agisoft Metashape.

The photogrammetric software reconstructs the scene by identifying the same details in many overlapping photos. Knowing where each detail appears in every image, it solves together the camera positions, the optical parameters of the lens, and the three-dimensional coordinates of the points. This step is called bundle adjustment and minimizes the overall error with which the 3D points, reprojected into the images, fall on the observed positions (Over et al., 2021). The ground control points anchor this reconstruction, in itself in an arbitrary system, to the real coordinates of the terrain. It is precisely in this balancing between image observations and ground markers that the choice of disabling the camera positions, described above, comes into play.

2.5.2 Rationale for the surface-level comparison

The accuracy assessment carried out in the previous paragraphs (§§ 2.3 and 3.1) concerns the corrected GPS coordinates at the control and check points, that is, a discrete set of positions. This positional accuracy is a necessary but not sufficient condition to establish the geometric fidelity of the reconstructed surface, which is the quantity truly relevant for the analysis of multitemporal variations. The GCP coordinates in fact enter the photogrammetric bundle adjustment. This propagates their errors into the geometry of the model in ways that the point-wise residuals on the check points do not fully capture (Liu et al., 2022; Zhong et al., 2025). In particular, a spatially structured error in the GCP elevations - such as the temporal drift of ΔZ documented in § 3.1.3 - can induce a systematic tilt of the reconstructed surface that the point-wise check does not reveal.

Since the research question concerns the suitability of the models for the analysis of surface variations over time, the accuracy must be evaluated at the surface level and expressed with respect to a detection threshold. The comparison between surfaces therefore has two objectives: (I) to quantify the geometric discrepancy between each GPS-georeferenced surface and the Total Station reference one; (II) to estimate the minimum level of detection (minLoD) allowed by the accuracy obtained, to be compared with the erosion expected on the tell. The first objective answers the question "how accurate is the low-cost model with respect to the gold standard". The second translates that accuracy into an operational criterion of detectability.

2.5.3 Dense point-to-point comparison: the M3C2 algorithm

Two three-dimensional surfaces can be compared with methods of different sophistication: from the simple elevation difference between rasters to the cloud-to-cloud distance along the direction of the nearest point. These elementary approaches are, however, sensitive to the spatially variable density and roughness and to planimetric misalignment - endemic conditions of the dense clouds produced by

Structure-from-Motion on irregular topographies such as a tell. To overcome them, the Multiscale Model-to-Model Cloud Comparison (M3C2; Lague et al., 2013) algorithm is adopted. It measures the distance between two clouds along the local normal to the surface rather than along a fixed direction.

M3C2 works at two scales and in two steps. Around each computation point (core point) it considers the neighbouring points within a sphere of radius D/2 - where D is the normal scale - and fits a plane to them. The perpendicular to the plane gives the local orientation of the surface (the normal), while the dispersion of the points around the plane measures its roughness. It then places a thin cylinder of radius d/2 - the projection scale - along that direction and identifies where the points of the two clouds fall within it. The mean position of each cloud provides two points, and their signed distance, measured along the normal and not along the vertical, is the local variation of the surface.

Each mean position has its own uncertainty, so M3C2 associates a confidence interval with each measurement: the smallest variation that can be trusted at 95%, that is, the local level of detection. With n_1 and n_2 points in the two cylinder segments and roughness σ_1 and σ_2, it is given by (Lague et al., 2013, p. 15, Eq. 1):

$$\mathrm{LoD}_{95\%} = 1.96 \cdot \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}} + \mathrm{reg}$$

Formula 2.8. M3C2 local level of detection (LoD 95%)

where reg is the co-registration error between the two clouds. A variation greater than the threshold is statistically significant, one smaller than it is not. The threshold therefore depends on the roughness of each cloud, on the number of points in the cylinder, and on the user-defined co-registration error. It is this formulation that makes the method robust to the variations of density and roughness typical of SfM dense clouds. The error budget can be made fully explicit by propagating the positional uncertainty point by point (Winiwarter et al., 2021); here its spirit is adopted through the co-registration term in the computation of the threshold.

Measuring along the normal, and not along the vertical, is what makes it possible to treat slopes and vertical walls correctly without any gridding, reducing the sensitivity of the comparison to planimetric misalignment on inclined surfaces. On flat terrain, imposing a vertical normal, M3C2 reduces exactly to the DoD (Lague et al., 2013). The user must set the two scales. The normal (D) must be chosen large enough not to be affected by roughness: the authors show that the orientation error becomes negligible when D is about 20-25 times the roughness at that scale (Lague et al., 2013, p. 15). The projection scale (d), instead, must average a sufficient number of points without erasing the real details.

In operational terms, before the computation it is verified that the two clouds share the same Global Shift, necessary for the correct superimposition of the UTM coordinates. The normal scale is not estimated automatically but set at about 20-25 times the mean point spacing of the dense cloud, so as to estimate the orientation on a neighbourhood stable with respect to the local roughness. The projection scale and the cylinder depth are dimensioned accordingly. The co-registration error used in the threshold derives from the accuracy of the control points of the two models. M3C2 is applied in CloudCompare by comparing the cloud to be evaluated (GPS-georeferenced scenario) with the reference cloud (Total Station scenario, S1). The latter is taken as the reference and the GPS cloud as the cloud of the core points. In this way each evaluation point corresponds to an actual position of the GPS surface, and the sign of the distance expresses the deviation from the gold standard.

2.5.4 Decomposition of the systematic error component

The total geometric discrepancy between two co-georeferenced surfaces decomposes into two parts. One is a systematic component - a mean shift and a rigid tilt that affect the entire surface in an ordered way - and the other is a random residual, made of small point-to-point oscillations. The systematic component is, by construction, a rigid-body transformation (a rotation and a translation). It is estimated in closed form with the algorithm based on the singular value decomposition of Eggert et al. (1997, pp. 273-274). In practice the same transformation is obtained in CloudCompare with a fine ICP registration (Iterative Closest Point; Besl & McKay, 1992; Lucks et al., 2024) between the two clouds, disabling the scale factor to impose a purely rigid model. The angle and the axis of rotation extracted from the transformation matrix directly measure the residual systematic tilt between the two surfaces. In the present dataset a component of this type is expected on a physical basis. The correction adopted in the main correction scenario removes the mean bias of ΔZ but cannot compensate for the spatial gradient produced by the temporal drift of ΔZ (§ 3.1.3). This gradient therefore remains as a residual tilt of the GPS surface with respect to the Total Station reference.

The treatment of this component depends on the analytical objective and must be kept explicit, so as not to confuse the accuracy assessment with change detection. To measure the absolute georeferencing accuracy - the comparison conducted here between scenarios of the same acquisition campaign - the tilt is retained and reported as part of the result. It is precisely the error we want to quantify, and it measures the geometric fidelity of the low-cost workflow. For the future multitemporal application, on the contrary, the systematic component shared by two epochs georeferenced on the same fixed points must be removed (co-registration), so that the residual reflects the real change. In that case the rigid transformation estimated above acts as a detrending operator. It is essential to note that it is the random residual obtained after this removal - and not the raw discrepancy - that provides the elevation uncertainty used in the detection threshold.

2.5.5 DEM of Difference and minimum level of detection

The comparison between digital elevation models of different epochs or configurations poses a fundamental interpretive problem: distinguishing the real variations of the surface from the combined noise of the two surveys. The minimum level of detection was introduced in the geomorphological monitoring of topographic variations from repeated surveys (Wheaton et al., 2010). It formalizes this distinction, setting a statistical threshold below which an observed difference cannot be confidently attributed to a real change, but falls within the combined uncertainty of the survey. The present work adopts this framework both for the computation of the DEM of Difference and for the subsequent volumetric estimate.

The DEM of Difference is obtained in CloudCompare. The dense clouds of the scenarios, already aligned, are first rasterized into regular-grid elevation models (0.10 m step, cell elevation equal to the mean of the points, same grid for the two scenarios being compared). The cell-by-cell difference and the corresponding volume are then computed with the 2.5D volumetric comparison tool, which sums the product of the elevation difference of each cell and its area (Wheaton et al., 2010, p. 138). The systematic component - the tilt - is already removed upstream by the rigid ICP registration (§ 2.5.4), so that the residual difference reflects only the random discrepancy between the surfaces. The statistical detection threshold, which has the dimensions of a length (Wheaton et al., 2010, p. 140), is computed as:

$$\mathrm{minLoD} = 1.96 \cdot \mathrm{NMAD}$$

Formula 2.9. DoD minimum level of detection (minLoD)

where NMAD is the robust estimate of the dispersion of the residual and the factor 1.96 corresponds to 95% confidence (Höhle & Höhle, 2009, p. 400). An elevation difference smaller than the minLoD cannot be confidently attributed to a real change, but falls within the uncertainty of the comparison. This threshold expresses the deviation of a single low-cost model from the reference. Its propagation between two epochs, proper to multitemporal monitoring, is beyond the scope of the present work.

Since the distribution of the residual is not necessarily Gaussian, the dispersion is estimated with robust statistics - the median and the Normalized Median Absolute Deviation (NMAD; Höhle & Höhle, 2009, p. 400) - consistently with the planimetric error of § 2.3.1. The classical standard deviation is not used, because the systematic component and the outliers would inflate it. A complementary observation concerns the sensitivity of the vertical difference to the slope: a residual planimetric error, denoted by e, on a surface of slope α produces an apparent elevation difference of about:

$$\Delta z_{\mathrm{app}} \approx e \cdot \tan\alpha$$

Formula 2.10. Apparent elevation difference from planimetric error on a slope so that the DoD overestimates the change on the steep flanks of the tell. Since M3C2 measures along the normal to the surface, it is comparatively insensitive to this effect. For this reason the two methods are reported jointly, so that their agreement increases confidence and their divergence localizes the slope-controlled artefacts. From the DoD the volumetric change is finally obtained, computed as the sum of the product of the elevation difference of each cell and its area:

$$V = \sum_i (\Delta Z_i \cdot A_{\text{cella}})$$

Formula 2.11. DoD volumetric change

with the positive and negative contributions kept distinct in order to separate accumulation and erosion (Wheaton et al., 2010, p. 138). In the present study the two models being compared come from the same acquisition campaign. The volume thus computed does not represent real erosion, but the apparent volume induced by the georeferencing difference alone and therefore measures the background volumetric noise of the workflow. The thresholded DoD excludes the cells below the minLoD from the sum, in order to isolate the real change (Wheaton et al., 2010, p. 140). Both this thresholded DoD and the propagation of the threshold between two epochs acquire meaning only in the real multitemporal comparison, which is beyond the scope of the present work.

One point is essential: the NMAD that enters the minLoD must be the "clean" random part of the residual, obtained after removing the systematic tilt (§ 2.5.4). Using the raw dispersion, inflated by the tilt, would give a meaningless threshold. It is the same logic as the confidence interval of the M3C2 (Lague et al., 2013, Eq. 1): in both cases a variation counts only if it exceeds the combined noise of the two surveys.

2.5.6 Design of the scenario comparison

The surface-level evaluation is articulated into three pairwise comparisons. The first compares the main correction scenario with the Total Station reference and measures the accuracy of the GPS correction estimated on all the available control points. The second compares the reduced control point configuration with the reference and measures the accuracy attainable in the operational configuration that future campaigns would adopt. The third compares the two correction scenarios with each other - main and reduced - and verifies, on the reconstructed surface, the equivalence already examined at the level of the corrected coordinates (§ 3.3).

The three comparisons are not independent: since the two correction scenarios differ by a constant translation, any two of the comparisons determine the third. This yields an internal consistency check: the deviation measured directly between the two correction scenarios must coincide, along the surface normal, with the constant translation that separates them (Table A7). It should be noted that this identity holds rigorously only for a quantity measured along a fixed common direction. The mean M3C2 deviations are measured along the local surface normal and referred to a distinct reference cloud in each pair. They are therefore orientation-dependent projections, not required to satisfy the arithmetic sum of the three pairwise means. The two accuracy comparisons are evaluated with both methods, M3C2 and DoD. The equivalence comparison, which reduces to a known constant translation, is verified only once. This design takes up the logic whereby the number and configuration of the high-precision control points govern the accuracy of the models derived from UAV (Martínez-Carricondo et al., 2018; Sanz-Ablanedo et al., 2018; Villanueva & Blanco, 2019). Here it is transferred from the photogrammetric bundle adjustment to the correction of the GPS coordinates.

3. Results

3.1 Structure of the GPS error and characterization of the positional residual

3.1.1 Descriptive statistics of the GPS-TS residuals

The residual between the consumer GPS coordinates and the Total Station ones was computed on the 23 common points. Its structure differs qualitatively between the planimetric components and the elevation one. Table 3.1 summarizes the descriptive statistics (Figures 3.1, 3.2, 3.3 and 3.4).

Variable	Mean (m)	Median (m)	SD (m)	Min (m)	Max (m)	RMSE (m)
ΔE	0.612	0.487	0.352	0.161	1.349	0.702
ΔN	0.433	0.461	0.349	−0.490	1.095	0.551
ΔZ	−14.016	−14.152	0.777	−15.439	−12.577	14.036
d2D	0.843	0.804	0.300	0.405	1.443	0.893
d3D	14.044	14.185	0.777	12.599	15.447	14.064

Table 3.1. Descriptive statistics of the consumer-GPS vs Total Station residuals (n = 23 points). RMSE = Root Mean Square Error; SD = Standard Deviation (Note: the d3D values are dominated by the vertical component ΔZ. The RMSE_Z value (14.036 m) reflects the systematic geoidal offset and not a random measurement error).

The East component (ΔE) is positive at all points, ranging from +0.161 to +1.349 m. All 23 points shift eastwards, so the GPS receiver has a systematic and directional bias with respect to the TS system. The mean (0.612 m) exceeds the median (0.487 m), a sign of a few high values that pull it upwards. The standard deviation (0.352 m) measures a non-negligible random noise superimposed on the bias. The SD/NMAD ratio is 0.929 (0.352 m against 0.379 m): the distribution has light tails and no relevant outlier in this component (Table 3.2).

Component	SD (m)	NMAD (m)	SD/NMAD ratio
ΔE	0.352	0.379	0.929
ΔN	0.349	0.214	1.627
ΔZ	0.777	0.772	1.007

Table 3.2. SD vs NMAD comparison for the components of the GPS-TS residuals. An SD/NMAD ratio > 1 indicates the presence of outliers that inflate the SD relative to the typical dispersion (Note: NMAD = Normalized Median Absolute Deviation = 1.4826 · median|xi − median(x)|. Höhle e Höhle (2009, p. 400).

The North component (ΔN) has the highest SD/NMAD ratio of the three, 1.627. Some outliers therefore inflate the classical estimate of the dispersion. The NMAD_N (0.214 m) equals 61% of the SD (0.349 m): the typical dispersion of the residuals in the North direction is 0.214 m, not 0.349 m. The main planimetric outliers are points 540 (ΔN = −0.490 m) and 530 (ΔN = +1.095 m). The elevation component (ΔZ) has a ratio of 1.007, almost unitary: the error in elevation is distributed almost perfectly normally around the mean bias, in line with the Shapiro-Wilk test ($p = 0.516$).

The planimetric residual d2D is not correlated with the distance from the nearest GCP, computed excluding the point itself (Table 3.3): $r = 0.057$, $p = 0.817$, effectively null (Figure 3.5). The planimetric error of the consumer GPS therefore has no systematic spatial structure and does not depend on the geometric coverage of the GCPs. The semivariograms of ΔE and ΔN confirm this, with irregular and structureless trends (Figures B1 and B2), consistent with an error of random character.

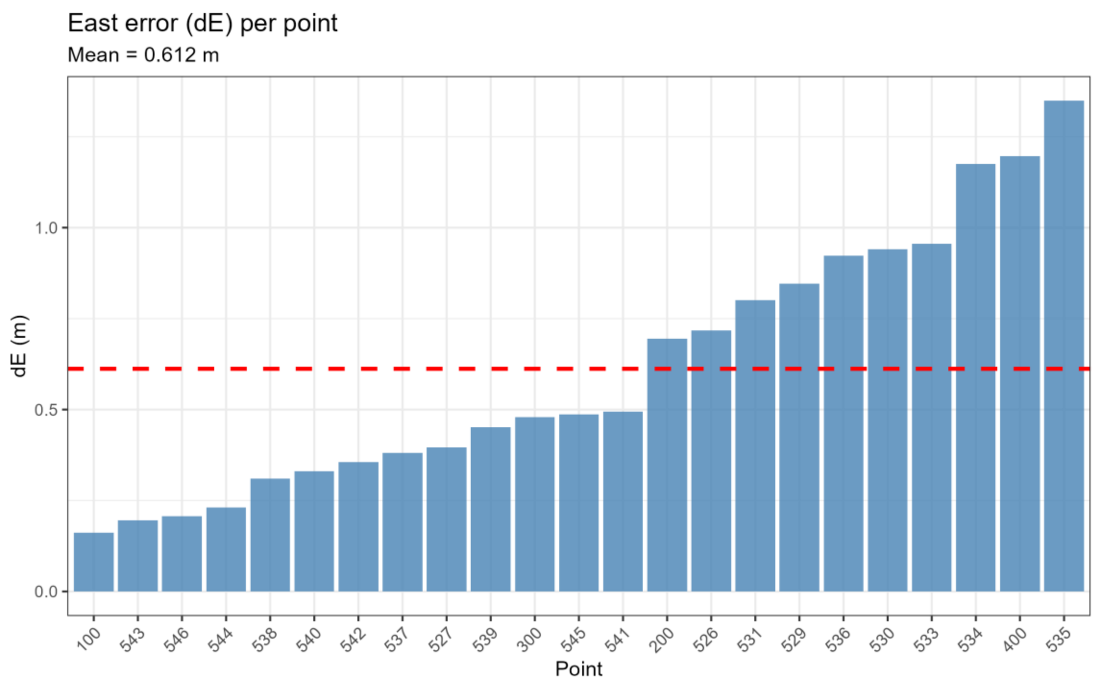

Figure 3.1. dE residuals per point (East component).

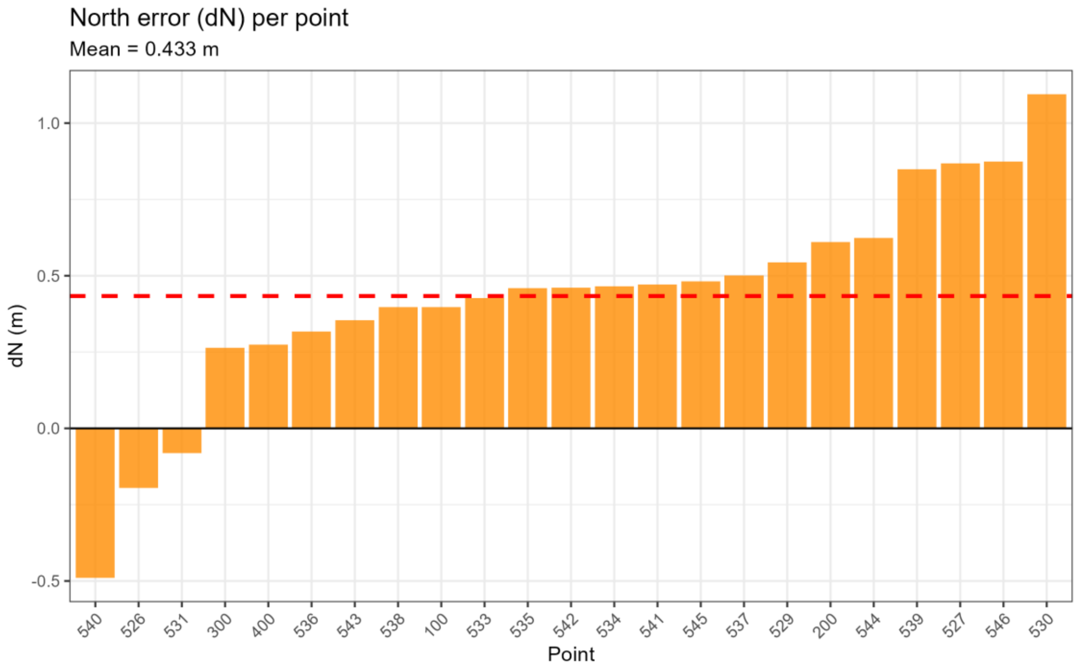

Figure 3.2. dN residuals per point (North component).

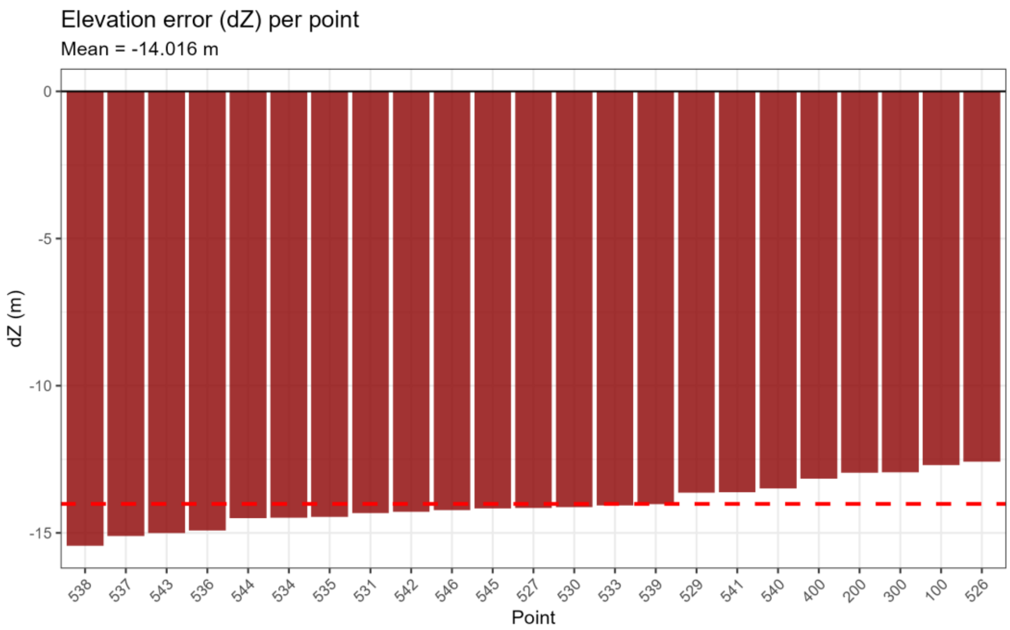

Figure 3.3. dZ residuals per point (vertical component).

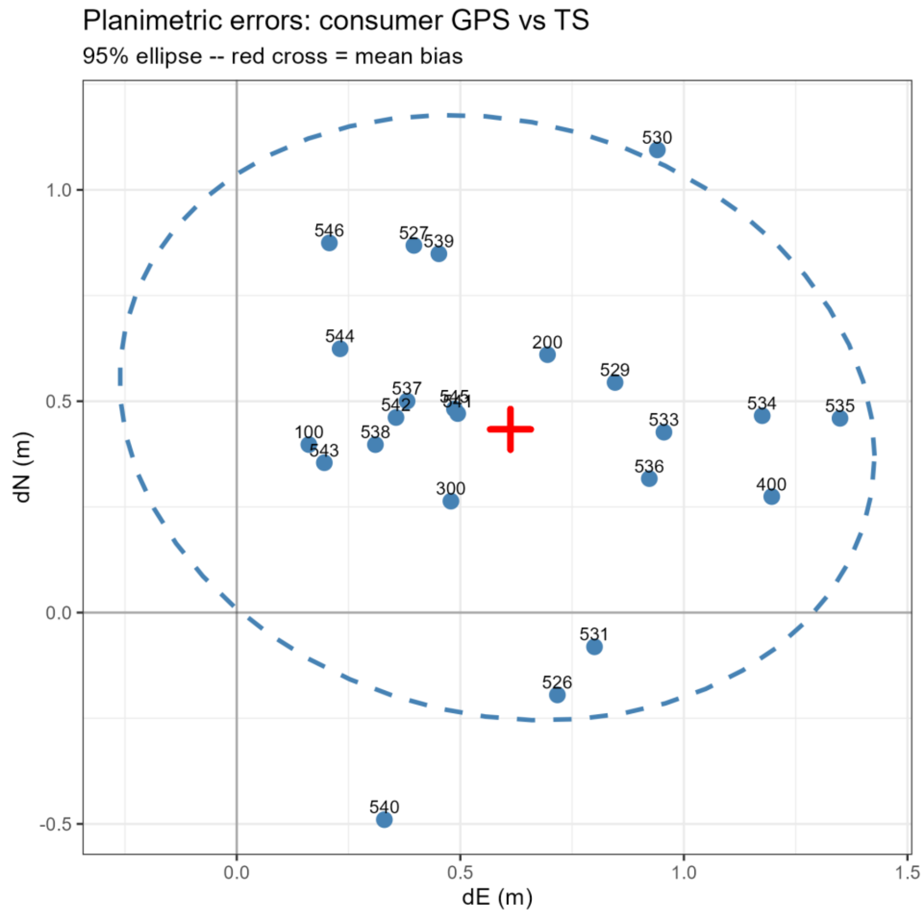

Figure 3.4. 95% confidence ellipse of the planimetric residuals.

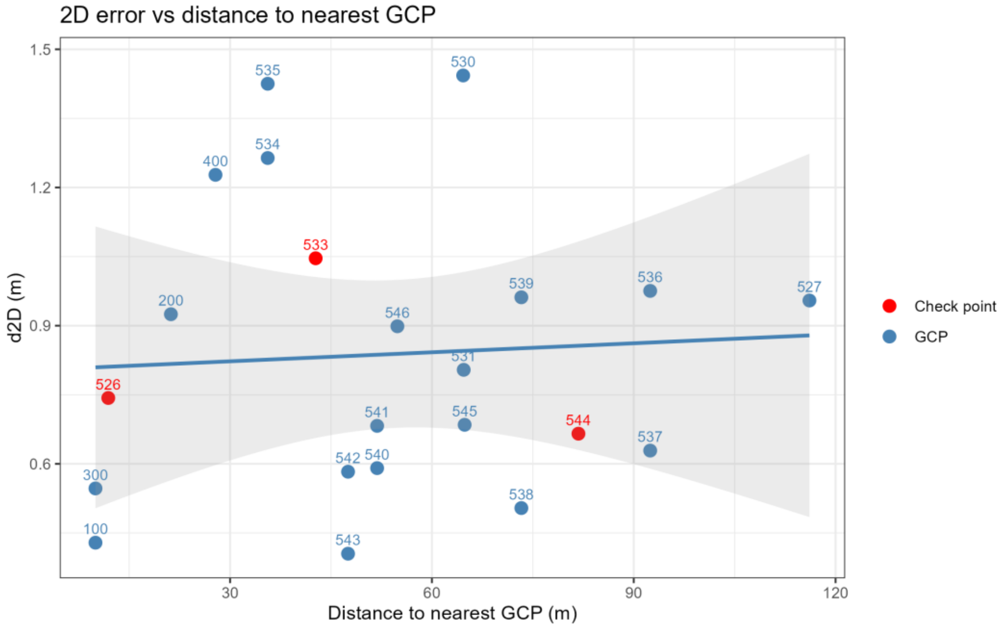

Figure 3.5. Planimetric (2D) error as a function of the distance to the nearest GCP.

relation	r	p
d2D_vs_distance_GCP	0.0571	0.8165

Table 3.3. Correlation between planimetric (2D) error and distance to the nearest GCP.

statistic	ratio
mean	1.00078358
sd	0.00508634
min	0.98205197
max	1.03091851

Table 3.4. Verification of the scale factor between GPS and Total Station.

3.1.2 Normality test

The Shapiro-Wilk test on the three components is compatible with normality in all cases (Figures 3.6, 3.7 and 3.8) (ΔE: $W = 0.924$, $p = 0.083$; ΔN: $W = 0.921$, $p = 0.071$; ΔZ: $W = 0.963$, $p = 0.516$). The borderline values of ΔE and ΔN ($p \approx 0.07$-0.08) indicate a slight asymmetry, due to the planimetric outliers of points 530 and 540. The p values remain above the 0.05 threshold but are close to the limit. With a small sample ($n = 23$) the power of the test is low, so scenario S5 adopts a robust estimator as a methodological safeguard.

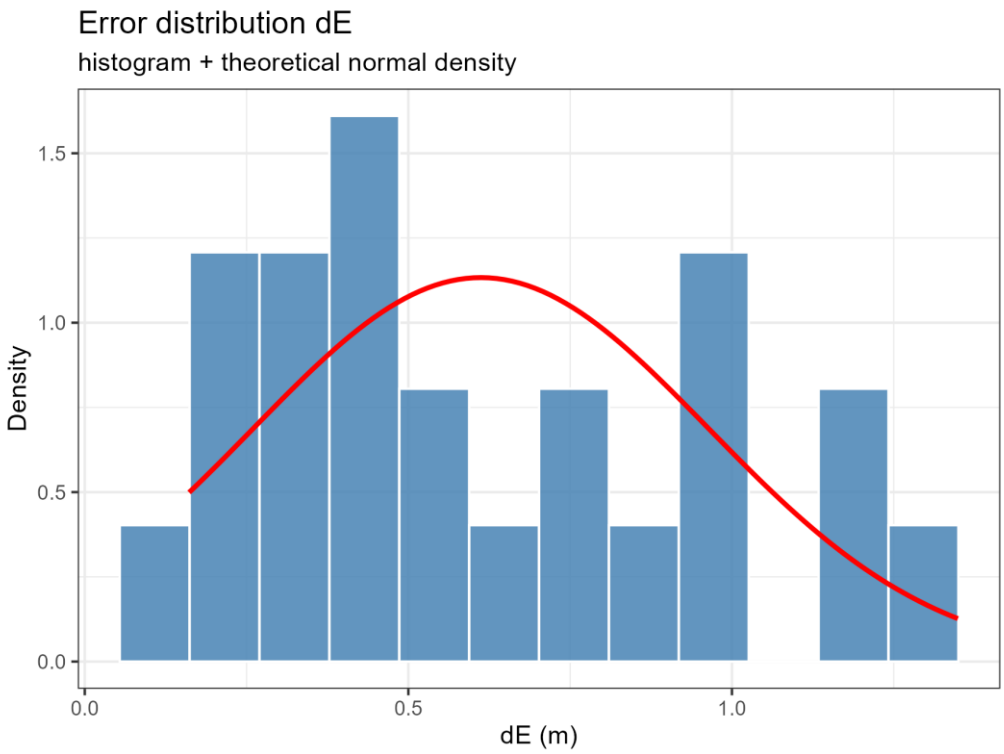

Figure 3.6. Normality histogram of the dE residual (East component).

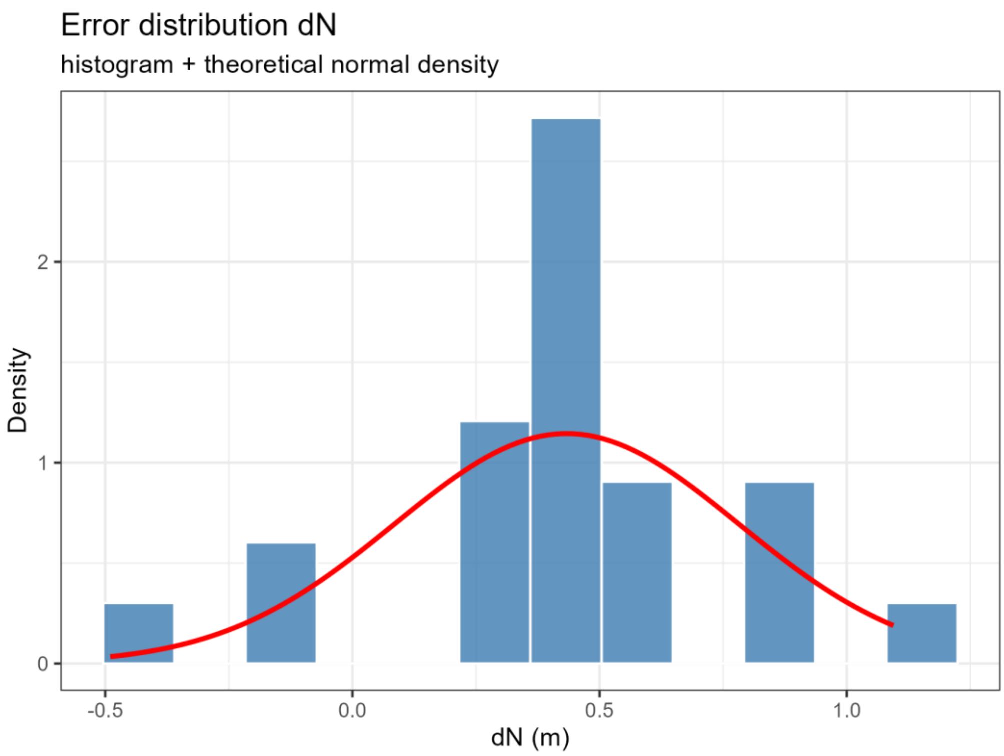

Figure 3.7. Normality histogram of the dN residual (North component).

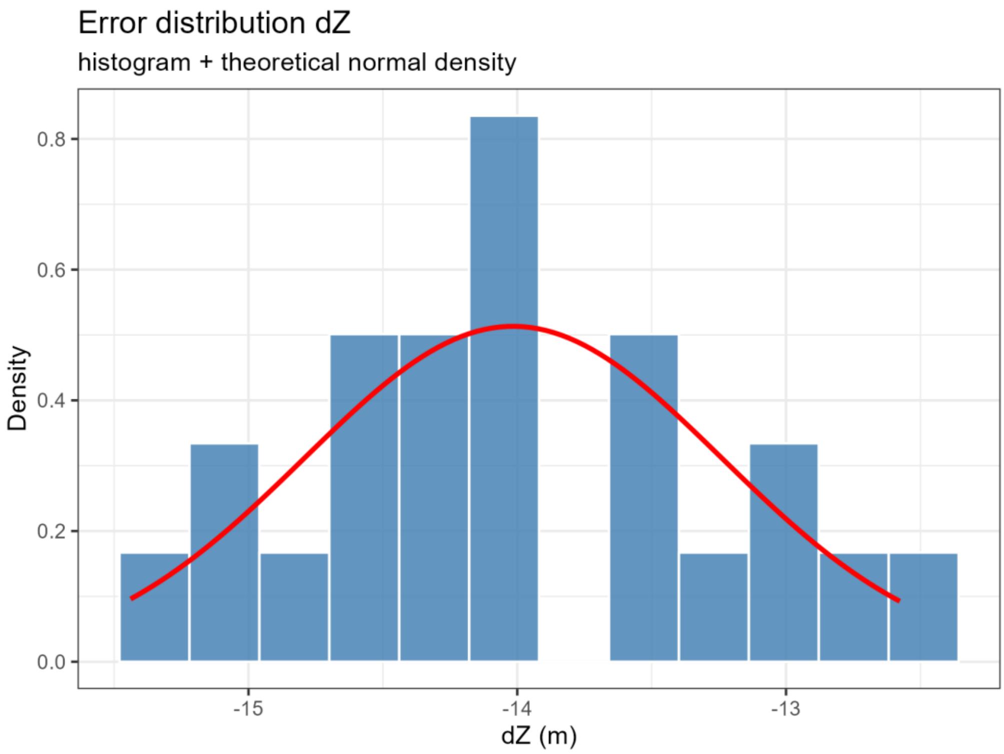

Figure 3.8. Normality histogram of the dZ residual (vertical component).

variable	W	p
dE	0.9244	0.0828
dN	0.9212	0.0708
dZ	0.9625	0.5161

Table 3.5. Shapiro-Wilk normality test on the residuals (E, N, Z).

3.1.3 Temporal correlation of the elevation error

The Pearson correlation between GPS acquisition time (UTC) and the elevation bias ΔZ is strong and highly significant: $r = -0.713$, $p = 0.000135$, 95% CI [−0.870, −0.426] (Figure 3.9). The Z bias becomes progressively more negative over the session, going from about −12.7 m at 06:16 UTC to about −15.1 m at 07:53 UTC, a drift of about 2.4 m in 1 hour and 37 minutes. The planimetric components do not correlate with time (ΔE: $r = +0.080$, $p = 0.717$; ΔN: $r = +0.014$, $p = 0.949$): the temporal drift concerns only the elevation. The semivariogram of ΔZ also grows almost monotonically with distance, from about 0.08 m^2 at 10-20 m to about 0.93 m^2 at 170-180 m (Figure B3). It is the spatial trace of the temporal drift: the points were acquired in spatial sequence, so the progressive variation of the Z bias produces a spatial correlation between the residuals of distant points.

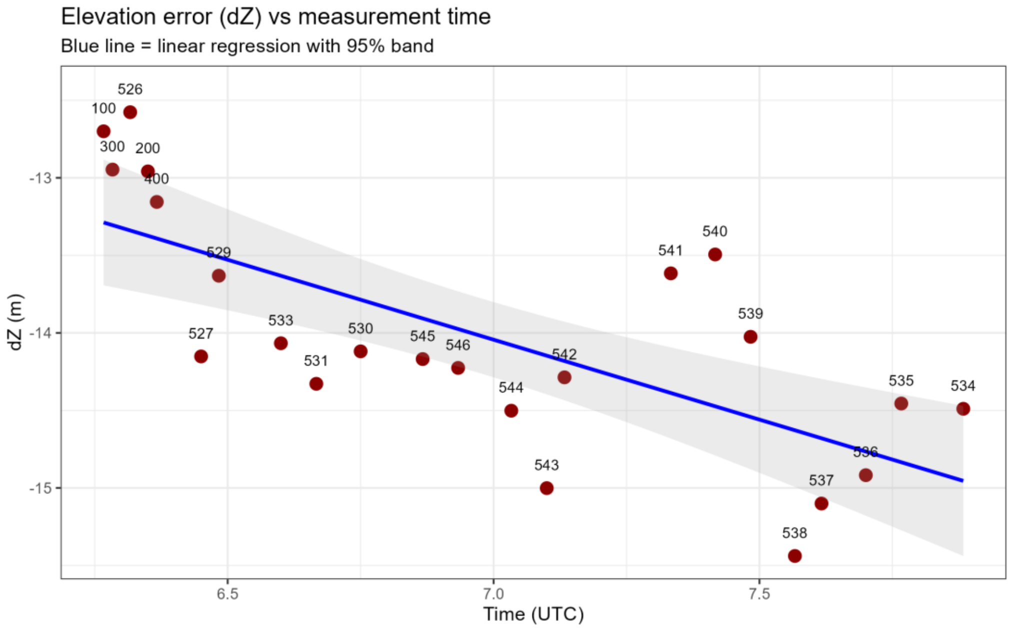

Figure 3.9. Scatter of the dZ residuals as a function of measurement time.

variable	r	p
dE	0.08	0.7166
dN	0.0141	0.949
dZ	-0.7128	1e-4

Table 3.6. Correlation between measurement time and residuals.

3.2 Performance of the coordinate correction models

The three models were evaluated on two criteria: the RMSE on the 19 GCPs (goodness of fit) and the mean RMSE on the 3 independent check points (real accuracy). Table 3.7 summarizes the results, which follow distinct patterns for the two criteria (Figure 3.10).

The correction parameters estimated on the 19 GCPs, which generate the corrected scenarios (§ 2.3.3), are the following. The pure translation (S3) is $t_E = -0.596$ m, $t_N = -0.451$ m, $t_Z = +14.083$ m (Table A1). The seven-parameter Helmert transformation (S4) gives negligible rotations ($|r_x| = 0.083°$, $|r_y| = 0.171°$, $|r_z| = 0.065°$), a scale factor of 0.99946 (−543 ppm), and translations E = −0.554 m, N = −0.472 m, Z = +14.077 m (Table A2). The preliminary analysis of the scale ratio between the two systems (premise of the model selection, § 2.3.2) returns a distribution with mean ≈ 1.00078, standard deviation ≈ 0.51%, and residuals between −1.80% and +3.09%. The scale variability is therefore dominated by the noise of the consumer GPS rather than by a real scale difference between the two systems.

Scenario	RMSE 2D GCP (m)	RMSE Z GCP (m)	Mean CP 3D (m)	Notes
S2 - Raw GPS	0.893	14.036	13.741	No correction
S3 - Translation 3p	0.493	0.752	0.862	Optimal method
S4 - Helmert 7p	0.459	0.702	0.911	Local coordinates
S5 - Helmert Huber	0.464	0.709	0.962	k=1.345 global

Table 3.7. Performance comparison of the GPS correction scenarios. GCP = Ground Control Points (n = 19); CP = independent check points (n = 3: ID 526, 533, 544). S1 (TS) is the reference: RMSE 2D = 0.070 m, RMSE Z = 0.007 m, Mean CP 3D = 0.141 m (Note: Mean CP 3D = mean of the 3D error on the three independent check points (ID 526, 533, 544). The values for S1 are: RMSE 2D GCP = 0.070 m, RMSE Z GCP = 0.007 m, Mean CP 3D = 0.141 m)

Scenario S2 (raw GPS, no correction) describes the typical conditions of a UAV survey georeferenced only with consumer GPS: RMSE 2D = 0.893 m on the GCPs and a mean 3D error of 13.741 m on the check points, dominated by the geoidal offset.

The pure three-parameter translation (S3) brings the RMSE 2D on the GCPs to 0.493 m and the RMSE Z to 0.752 m: −44.8% and −94.6% with respect to S2. The mean 3D error on the check points drops to 0.862 m. The individual check points vary greatly (complete residuals per scenario in Table A5): CP 533 (0.360 m) and CP 544 (0.582 m) have residuals in line with the noise expected after the correction of the mean bias, while CP 526 (1.643 m) is markedly higher. This residual reflects the temporal drift of the Z bias: the point was acquired at 06:19 UTC, when the bias was about −12.7 m against the corrected mean value of −14.083 m.

The seven-parameter Helmert transformation with local coordinates (S4) improves the RMSE 2D on the GCPs only slightly (0.459 m, −6.9% with respect to S3) and the RMSE Z (0.702 m, −6.6%), but on the check points the mean 3D rises to 0.911 m. CP 544 worsens greatly (from 0.582 m to 1.077 m), while CP 533 improves (from 0.360 m to 0.157 m).

Scenario S5 (Helmert with global Huber estimator) does not improve on S4: the mean 3D on the check points is 0.962 m (parameters in Table A3; Huber weights per point in Table A4).

Considering GCPs and check points together, the pure three-parameter translation (S3) gives the lowest mean 3D on the check points (0.862 m). The interpretation of these outcomes - non-separability of the Helmert parameters and the role of the robust estimator - is taken up in § 4.1.

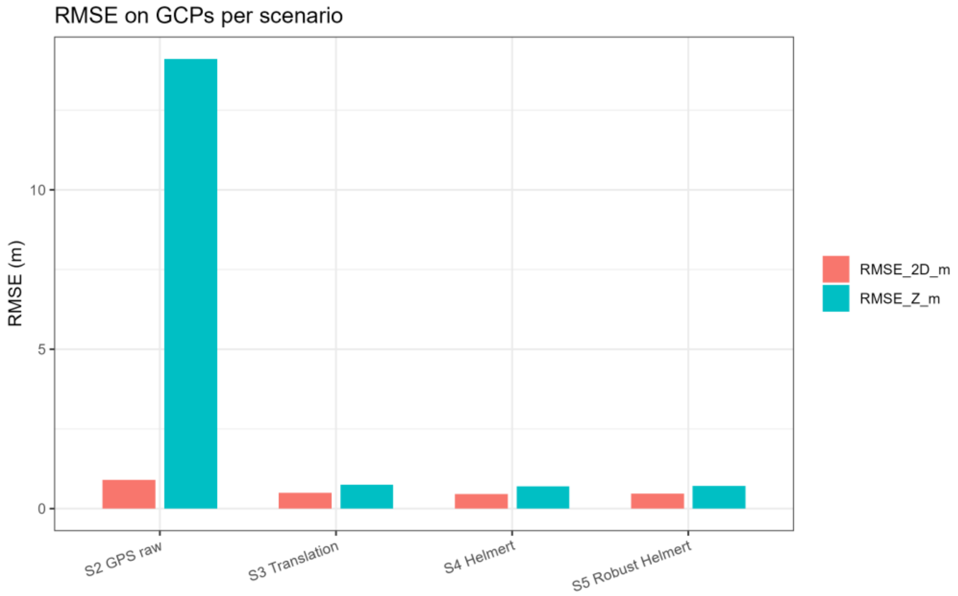

Figure 3.10. RMSE on the GCP for each georeferencing scenario.

3.3 Minimum number of reference points: results of the Monte Carlo analysis

The Monte Carlo analysis produced 8,500 realizations (17 values of n × 500 iterations each). Table 3.8 summarizes the results.

n GCP	Med 2D (m)	P05 2D (m)	P95 2D (m)	Med Z (m)	P05 Z (m)	P95 Z (m)
3	0.538	0.468	0.686	0.924	0.825	1.320
5	0.510	0.465	0.604	0.920	0.826	1.180
7	0.506	0.465	0.590	0.905	0.826	1.120
10	0.499	0.468	0.554	0.906	0.830	1.040
15	0.493	0.475	0.520	0.905	0.852	0.971
19	0.491	0.491	0.491	0.903	0.903	0.903

Table 3.8. Results of the Monte Carlo analysis: median and 5-95% percentiles of RMSE 2D and RMSE Z on the check points for selected values of n GCP (500 iterations per value of n, fixed seed set.seed(42) (Note: Med = median; P05 = 5th percentile; P95 = 95th percentile).

The median of the RMSE 2D drops from 0.538 m with n = 3 to 0.491 m with n = 19, in all only 0.047 m (8.7% relative). The curve has a clear knee around n = 7 GCPs. With 3 GCPs the median is 0.538 m, with 7 GCPs it is 0.506 m: relative to the uncorrected GPS (S2, RMSE 2D = 0.893 m), seven points already recover 96% of the planimetric correction attainable with all 19 GCPs. Beyond n = 7, each additional GCP improves the median by less than 2% (Figure 3.11).

The 5-95% confidence band narrows markedly as n increases: from 0.218 m with n = 3 GCPs (0.468-0.686 m) to 0.125 m with n = 7 (0.465-0.590 m). The main advantage of more GCPs is therefore not a better median, but a lower risk of poor results due to an unfavourable selection of the sample, in line with Villanueva and Blanco (2019, p. 173) and Zhong et al. (2025, p. 2016).

The median of the RMSE Z remains almost flat for every n (range 0.896-0.924 m; the lower bound of 0.896 m corresponds to an intermediate value of n not among the representative rows reported in Table 3.8): increasing the number of GCPs of the same type does not reduce the residual elevation error (Figure 3.12). The cause is the temporal drift of the Z bias of § 3.1.3. The error is not random but structured in time, and no number of static GCPs can correct a dynamic variation of the signal.

Overall, 7 Total Station GCPs are the optimal operational threshold for the proposed correction workflow: they balance the gain in planimetric accuracy, the lower variability between campaigns, and the cost of the survey with precision instrumentation. The recommended distribution includes perimeter points, at least one central point, and coverage of the available elevation range, following Martínez-Carricondo et al. (2018) for the optimal distribution of GCPs in UAV photogrammetry on terrain of variable morphology (Figure 3.13).

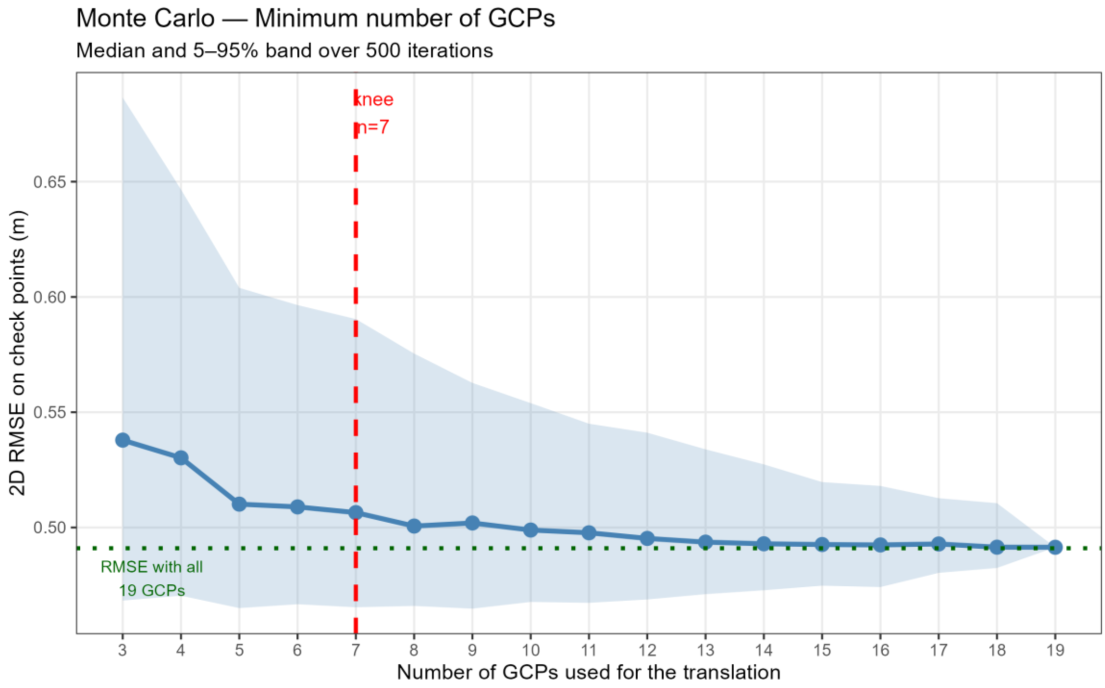

Figure 3.11. Monte Carlo simulation: distribution of the planimetric (2D) RMSE.

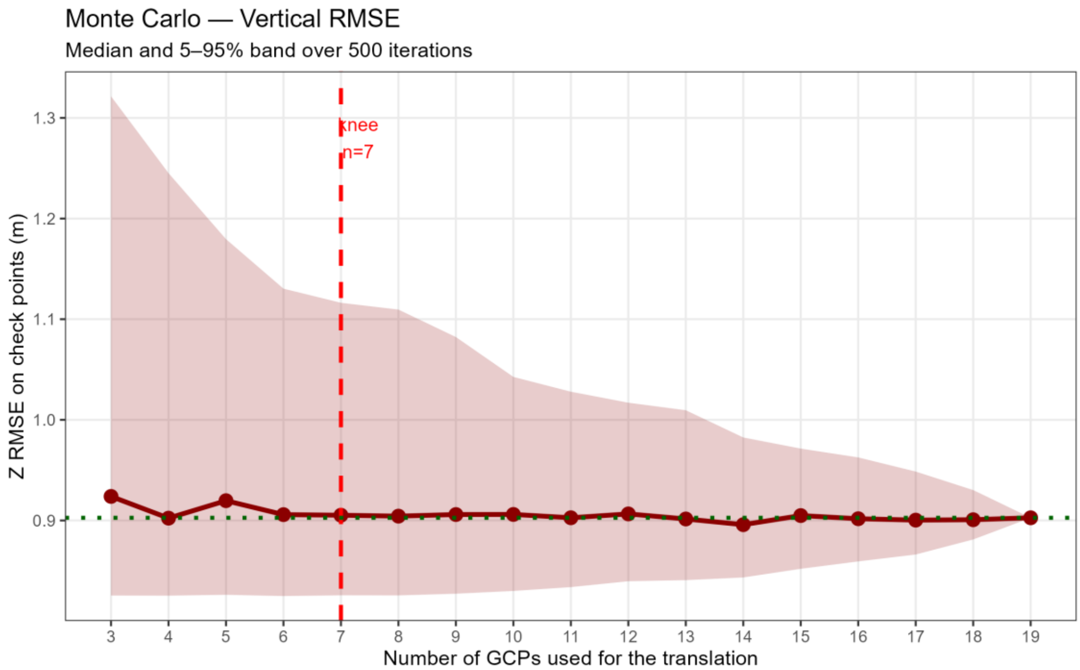

Figure 3.12. Monte Carlo simulation: distribution of the vertical (Z) RMSE.

Figure 3.13. Spatial distribution of the GCP in the study area.

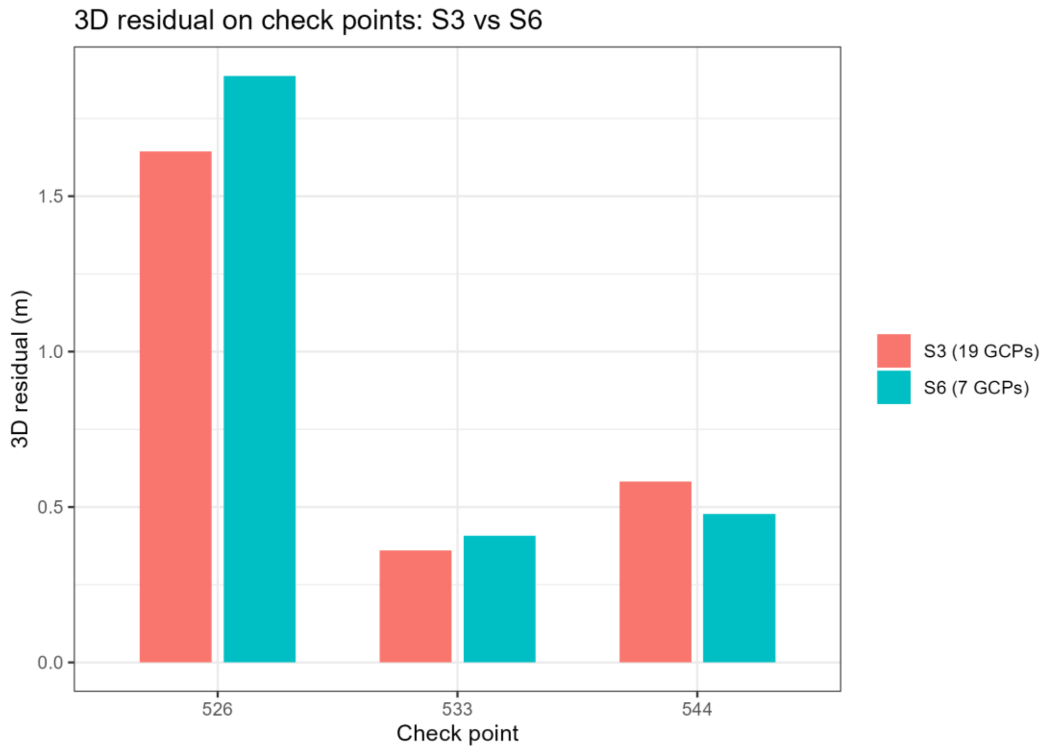

Figure 3.14. 3D residual on the check points: comparison between scenarios S3 and S6.

3.4 Analytical framework for surface-level accuracy assessment

The coordinate-level results of §§ 3.1-3.3 lead to two outcomes. With the pure translation (S3) the corrected GPS coordinates have a mean 3D error of 0.862 m on the independent check points (S6 residuals and comparison with S3 in Tables A8 and A9; Figure 3.14). The Monte Carlo analysis sets at seven GCPs the threshold beyond which each additional control point gives a negligible benefit. These outcomes describe the accuracy of the corrected coordinates, but not the geometric fidelity of the reconstructed surfaces - the quantity that matters for the analysis of multitemporal variations (§ 2.5.2). On this basis, three scenarios are carried through to the full photogrammetric reconstruction - the Total Station reference (S1), the correction estimated on all the available GCPs (S3), and the seven-GCP configuration (S6; parameters in Table A6) - while the others remain evaluated only at the coordinate level.

The surface-level comparison defined in § 2.5 translates the accuracy of the coordinates into a statement on the geometry of the models and on detectability. For each accuracy comparison (S3-S1 and S6-S1) the total discrepancy was decomposed into two parts: a systematic component - the residual tilt due to the temporal drift of ΔZ (§ 3.1.3), which the pure translation does not compensate - and a random residual. The residual, estimated with robust statistics (Höhle & Höhle, 2009), provides the elevation uncertainty that enters the minimum level of detection minLoD = 1.96 · NMAD (Höhle & Höhle, 2009, p. 400). Comparing this threshold with the erosion expected on the tell shows whether the low-cost workflow is accurate enough for the analysis of multitemporal surface variations. The quantitative surface results are reported in § 3.4.2.

3.4.1 Accuracy of the reconstructed models

The sparse cloud cleaning workflow (§ 2.5.1) brings the mean reprojection error below 0.5 pixels. Two different accuracy metrics, treated together so far, must now be distinguished. The 3D error of 0.862 m reported in § 3.2 is a coordinate-level residual: it measures the difference between corrected GPS coordinates and reference coordinates at the check points, computed in R independently of the photogrammetric processing. The model-level accuracy is the RMSE on the check points after the bundle adjustment in Metashape - with the camera positions disabled from the optimization (§ 2.5.1) - and measures the geometric fidelity of the reconstructed 3D model. For scenario S3 it is about 2.63 m on the control points and 3.87 m on the independent check points: a value dominated by the elevation and one order of magnitude greater than the coordinate residual.

The gap is not a worsening of accuracy, but the direct and expected consequence of a choice: assigning the GPS markers a deliberately loose elevation accuracy (Z = 20 m, § 2.2). By leaving the absolute elevation weakly constrained, the bundle adjustment is prevented from tilting or deforming the surface to absorb the residual elevation bias. The price is a vertical residual of a few metres on the model, concentrated entirely in the Z component, while the planimetry remains accurate to the metre. In scenario S1 the Total Station markers (0.005 m) prevail over the cameras anyway, so the model accuracy remains governed by the precision markers. Distinguishing coordinate accuracy and model accuracy is essential to correctly read the surface results that follow

(§ 2.5.2), where the overall geometric discrepancy is measured directly on the reconstructed surfaces.

3.4.2 Results of the surface comparison

The comparison between the optimal low-cost scenario (S3, GPS with translation on 19 GCPs) and the Total Station reference (S1) was performed with the M3C2 algorithm on the definitive dense clouds, processed with the cameras disabled from the optimization and with identical parameters (Medium quality, Mild depth filtering; about 91.4 million points per cloud). The two clouds share the same Global Shift (−312000, −4066000, 0) and the reference system EPSG:32638. The M3C2 parameters adopted are: normal scale of 0.251 m in diameter, projection scale of 0.502 m, maximum depth of 34.2 m, evaluation points on a subsample at 0.25 m, and null co-registration error. S3 is the cloud to be compared and S1 the reference, so the sign of the distance expresses the deviation of the GPS surface from the gold standard.

Pair	Mean bias μ (m)	ICP tilt	Residual σ (m)	Residual NMAD (m)	minLoD (m)
S3-S1	+0.168	1.478°	0.246	0.266	0.52
S6-S1	−0.538	1.479°	0.240	0.223	0.44
S3-S6	+0.355	-	0.133	-	-

Table 3.9. Decomposition of the M3C2 surface comparison for the three pairs: mean bias, systematic tilt estimated from the ICP registration, residual dispersion (σ, NMAD) and minimum level of detection (minLoD = 1.96·NMAD). (Note. For the S3-S1 and S6-S1 pairs the mean bias is the mean of the raw M3C2 distance (mean deviation from the Total Station reference), while σ and NMAD are computed on the residual after the ICP registration. For the S3-S6 pair (internal consistency, in the absence of an absolute reference) the tilt shared by the two scenarios cancels in the direct comparison: μ and σ are measured on the raw comparison, without registration).

The raw M3C2, computed before removing the systematic component, gives a mean deviation close to zero (μ = 0.168 m; median 0.095 m) but a high dispersion (σ = 3.402 m; NMAD = 3.498 m), with 95% of the distances between −6.07 and +6.84 m (Figure 3.16). The spatial pattern of this systematic component is shown in Figure 3.20.

The fine rigid-body ICP registration (§ 2.5.4) recognizes in this dispersion a systematic tilt of 1.478° around an almost horizontal axis (0.712; 0.699; 0.065). The direction of the axis is consistent with the temporal drift of ΔZ described in § 3.1.3 (r = −0.713). The value measured on the dense clouds (1.478°) coincides within 0.01° with that estimated on the pilot comparison of the tie points (1.487°).

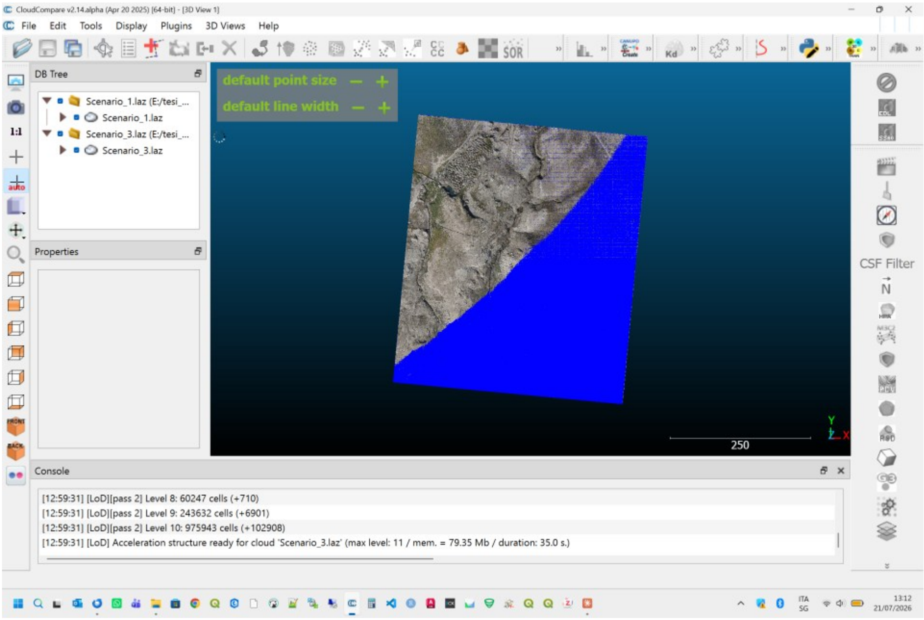

Figure 3.15. Superimposition of the two point clouds (S3 vs S1) showing the systematic tilt before ICP registration (©ReLand).

Once the rigid systematic component is removed, the M3C2 on the residual gives a null mean deviation (μ = 0.002 m) and a dispersion of 0.246 m (NMAD = 0.266 m; Figures 3.19 and 3.21). Here the NMAD slightly exceeds the standard deviation (0.266 m against 0.246 m). This is consistent with a residual distribution whose tails are lighter than Gaussian - that is, free of the outliers that would instead inflate the standard deviation above the NMAD - so the robust minLoD adopted here is, if anything, marginally conservative. The same relation holds for the corresponding DoD residual (§ 3.4.3, NMAD 0.249 m against σ 0.233 m). The systematic tilt alone therefore explains 99.5% of the raw variability. Applying the formula of Wheaton et al. (2010) to the residual, the minimum level of detection is of the order of 0.5 m (minLoD = 1.96·NMAD, with NMAD estimated robustly; Höhle & Höhle, 2009, p. 400).

The S6-S1 pair shows the same pattern as S3-S1. The ICP registration identifies a tilt of 1.479° around an almost horizontal axis, coinciding with that of S3 within 0.001°. Here too the systematic component explains 99.5% of the raw variance, and the residual reduces to σ = 0.240 m (NMAD = 0.223 m), values practically identical to those of S3-S1. The difference with respect to S3 is concentrated in the mean bias: with seven control points the mean deviation from the reference is −0.538 m, against the 0.168 m of S3 (Table 3.9) (Figures 3.17 and 3.18).

The S3-S6 pair evaluates the internal consistency between the two GPS corrections (nineteen and seven control points). Since neither scenario constitutes the absolute reference, the choice of the reference model is conventional and determines only the sign: S3 (optimal correction with nineteen points) was set as the model to be compared and S6 (reduced correction with seven points) as the reference, so the distance expresses the deviation of the reduced scenario from the optimal one. The two scenarios share the same tilt, which cancels in the direct comparison, so the ICP registration is not necessary: the raw comparison already presents a contained dispersion (σ = 0.133 m) around a mean deviation of 0.355 m. This deviation corresponds to the bias difference between the two scenarios and is consistent with the translation vector that separates them (Table A7); the reduced dispersion indicates that the two surfaces are geometrically almost identical in shape (Table 3.9). This identity is expected to hold only for a comparison measured along a fixed common direction. The difference between the mean M3C2 biases of the S3-S1 and S6-S1 pairs (0.706 m) is not required to equal the direct S3-S6 deviation (0.355 m), because the M3C2 distance is measured along the local surface normal and each pair is referred to a different reference cloud, so the mean bias is an orientation-dependent projection rather than a scalar offset. The relevant consistency is therefore that the direct S3-S6 deviation reproduces, along the surface normal, the constant translation vector that separates the two corrections (Table A7), and not that it equals the difference of the M3C2 mean biases of the other two pairs. The interpretation of these comparisons - role of the tilt, noise floor, and detectability - is taken up in § 4.3 and § 4.4.

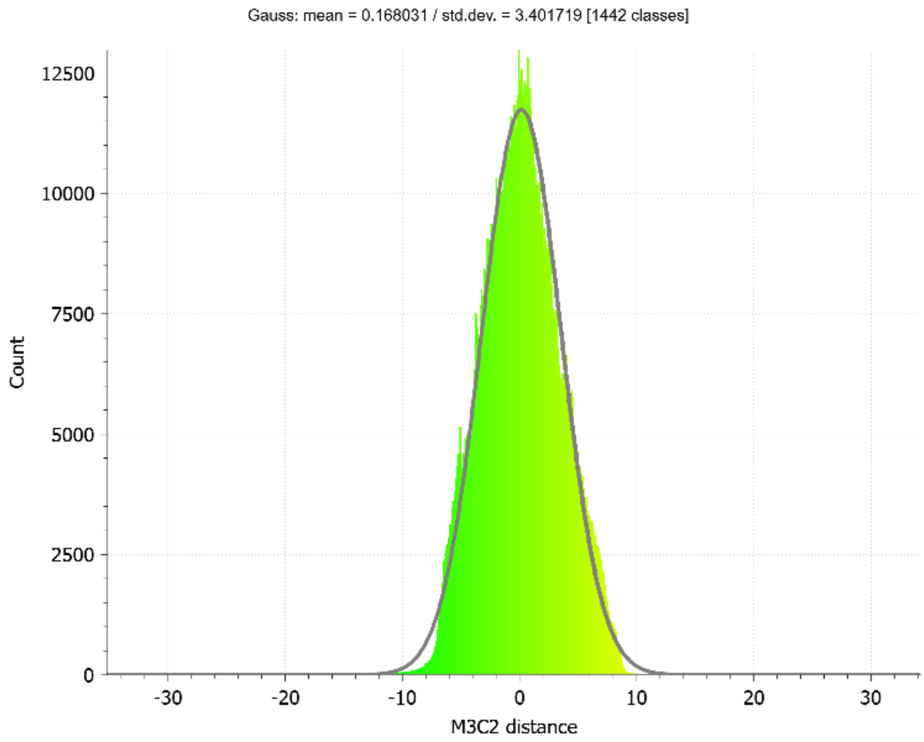

Figure 3.16. Histogram of the M3C2 distance, pair S3-S1 (raw).

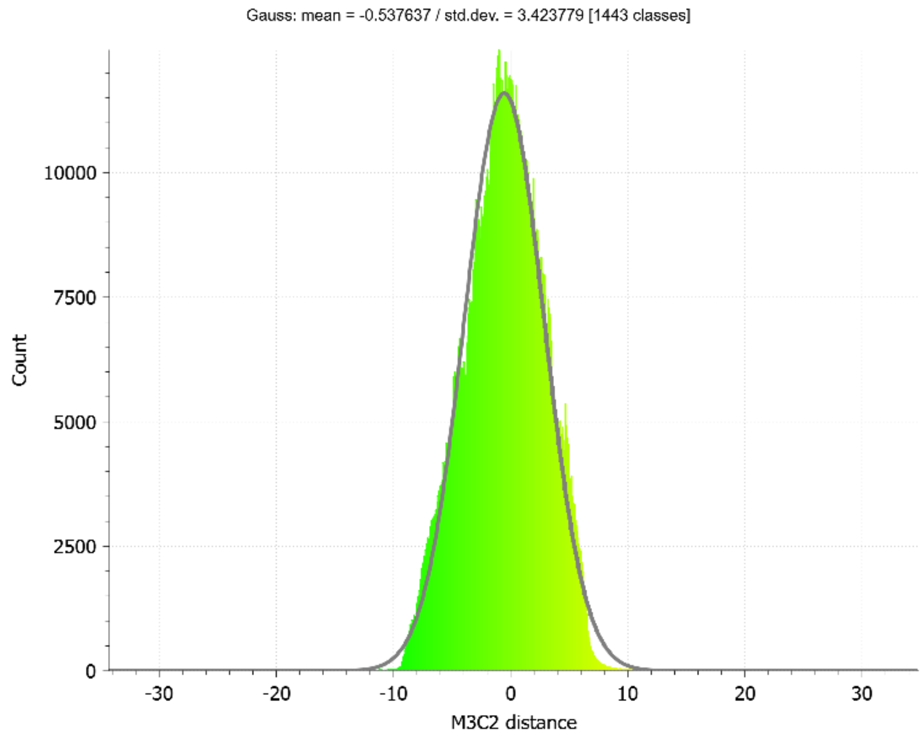

Figure 3.17. Histogram of the M3C2 distance, pair S6-S1 (raw).

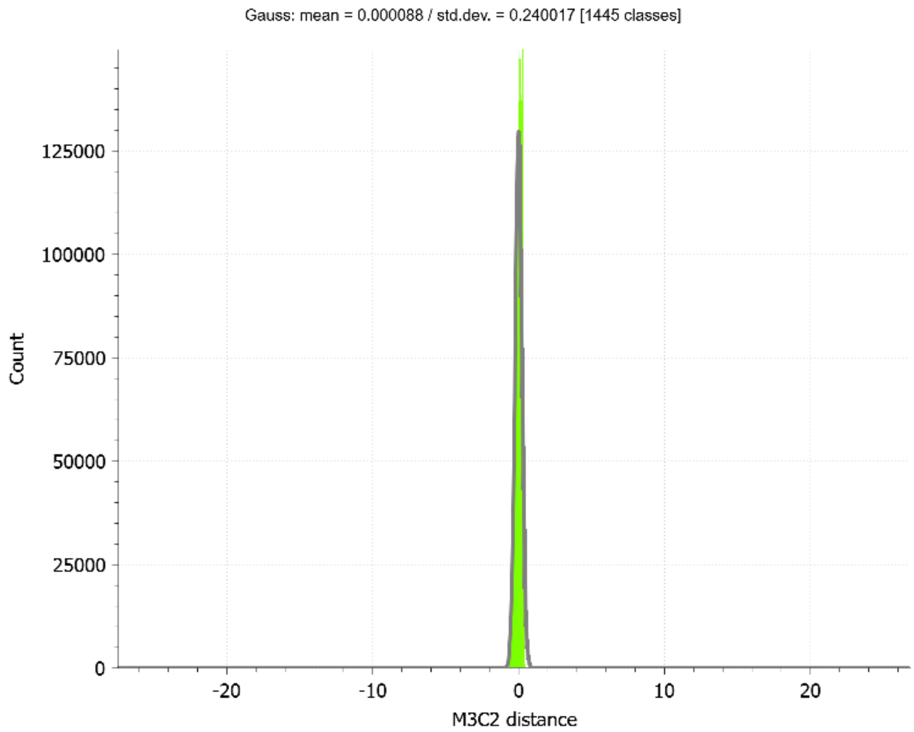

Figure 3.18. Histogram of the M3C2 distance, pair S6-S1 (post-ICP registration).

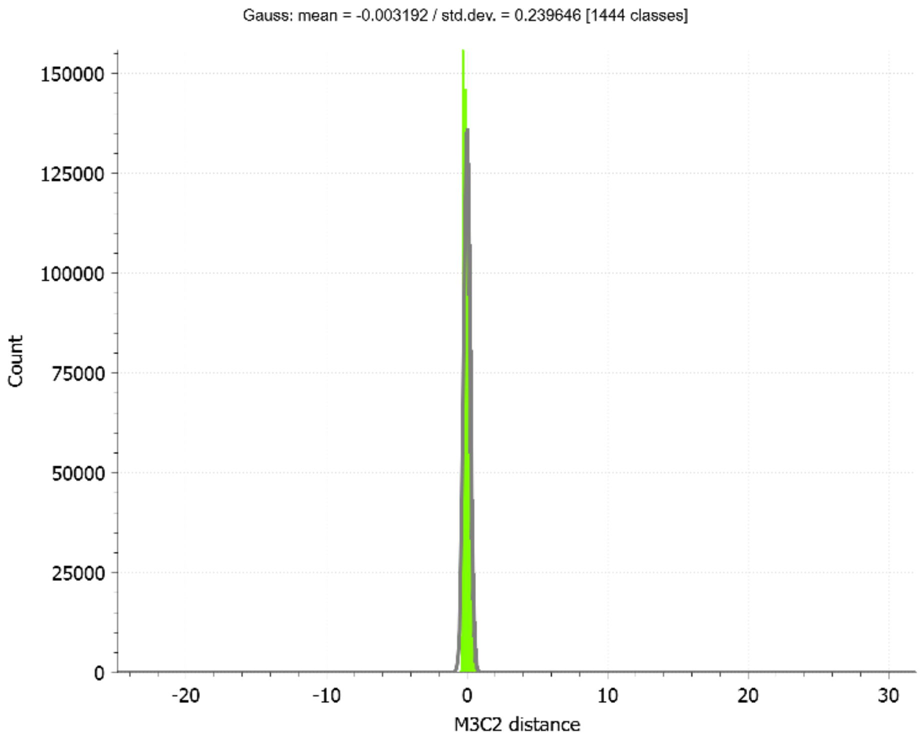

Figure 3.19. Histogram of the M3C2 distance, pair S3-S1 (post-ICP registration).

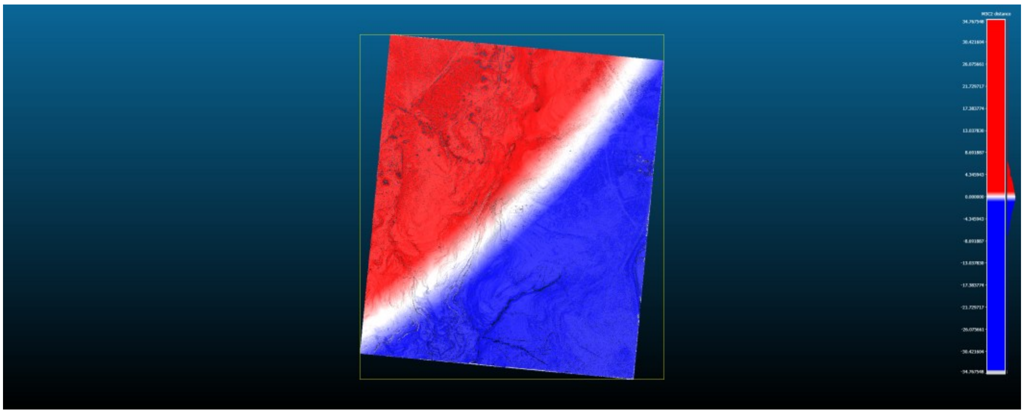

Figure 3.20. Colour-coded M3C2 distance on the surface, pair S3-S1 (raw): the systematic tilt (colours saturated at ±1 m).

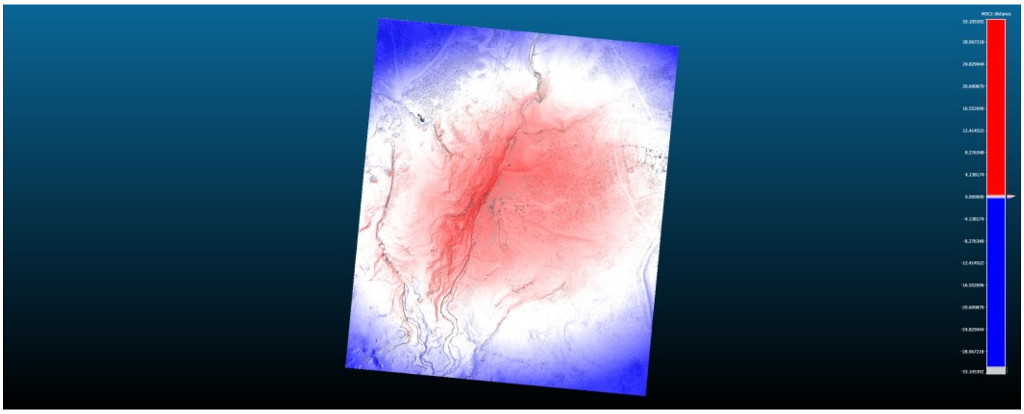

Figure 3.21. Colour-coded M3C2 distance on the surface, pair S3-S1 (post-ICP registration): the random residual after tilt removal (colours saturated at ±0.5 m).

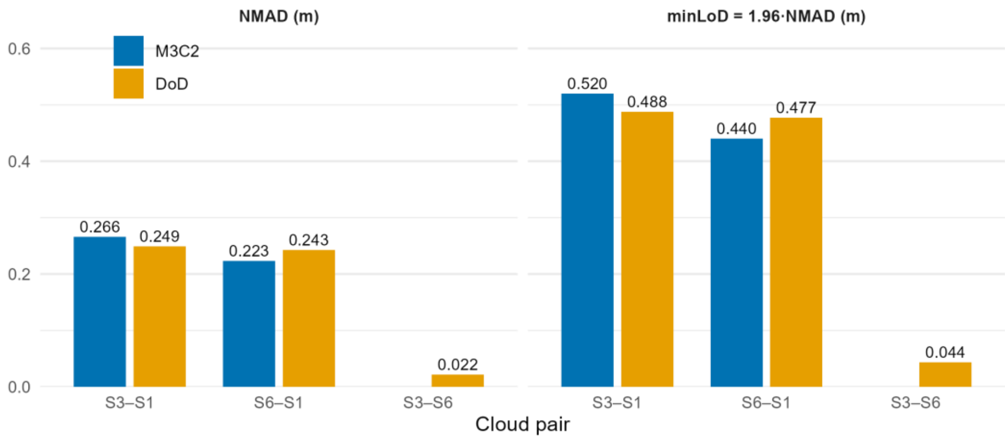

Figure 3.22. Comparison of NMAD and minLoD between the M3C2 and DoD methods.

pair	NMAD_M3C2_m	minLoD_M3C2_m	NMAD_DoD_m	minLoD_DoD_m
S3-S1	0.266	0.52	0.249	0.488
S6-S1	0.223	0.44	0.243	0.477
S3-S6	NA	NA	0.022	0.044

Table 3.10. Comparison of NMAD and minLoD between the M3C2 and DoD methods (cf. Table 3.9).

3.4.3 Results of the DoD (DEM of Difference)

The DEM of Difference confirms, independently, the results of the M3C2 comparison. The three comparisons are conducted on the 0.10 m grids of the post-ICP models, with a common-cell coverage of 99.9%. Table 3.11 reports, for each pair, the robust statistics of the residual, the minimum level of detection (minLoD = 1.96 · NMAD), and the apparent volume (Figures 3.24, 3.25 and 3.26).

Pair	median (m)	σ (m)	NMAD (m)	minLoD (m)	Net vol. (m^3)	Deposition (m^3)	Erosion (m^3)	Area (m^2)
S3-S1	−0.045	0.233	0.249	0.488	−1245.870	+18092.505	−19338.374	194767.896
S6-S1	−0.043	0.232	0.243	0.477	−559.258	+18311.175	−18870.433	194735.776
S3-S6	−0.004	0.052	0.022	0.044	−696.579	+2194.407	−2890.987	194845.056

Table 3.11. DoD results for each pair (post-ICP residual, 0.10 m grid). Robust statistics of the residual (median, σ, NMAD), minimum level of detection minLoD = 1.96 · NMAD, net apparent volume and deposition/erosion contributions, common area. Common cells: 99.9% in all pairs. Volumes are apparent (same epoch) and quantify the volumetric noise of the workflow, not real erosion. Sources: cell-by-cell difference and volume (Wheaton et al., 2010, p. 138); NMAD (Höhle & Höhle, 2009, p. 400, Eq. 2); minLoD = 1.96 · NMAD, with 1.96 = 95% (Höhle & Höhle, 2009, p. 400).

The random residual measured by the DoD (NMAD 0.249 m for S3-S1, 0.243 m for S6-S1) coincides with that of the M3C2 (§ 3.4.2, NMAD 0.266 and 0.223 m): two independent methods - one along the normal to the surface, the other on the vertical grid difference - converge on the same noise floor (Table 3.10; Figures 3.22 and 3.23) of about 0.24 m, from which a minLoD of the order of 0.5 m. This floor is identical in the two scenarios, at nineteen and at seven control points. The S3-S6 pair, which compares two corrections of the same flight, presents a much lower dispersion (NMAD 0.022 m) and gross deposition and erosion volumes about one order of magnitude smaller (Table

3.12) - while the net volume, being their difference, is reduced by roughly a factor of two - because the common tilt and bias cancel (DoD distributions for S6-S1 and S3-S6 in Figure B4 and Figure B5). The volumes of Table 3.11 are apparent: deriving from models of the same epoch, they measure the volumetric noise of the workflow, not a real erosion.

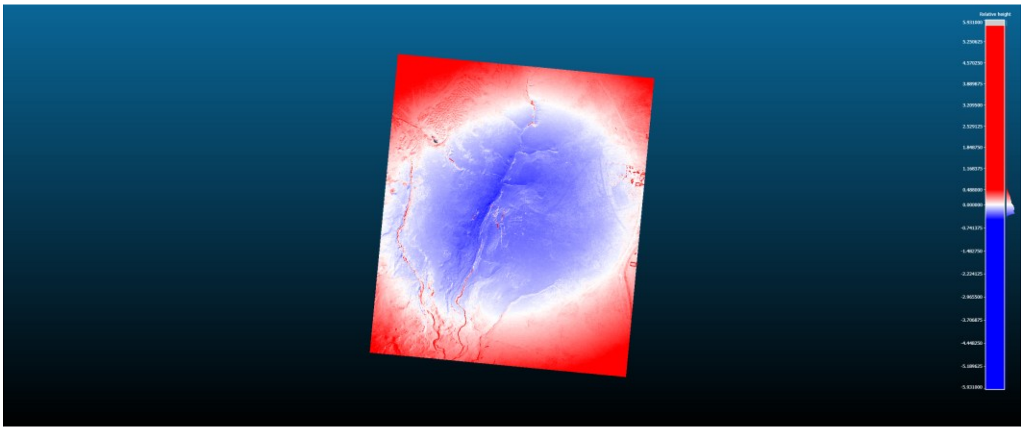

Figure 3.23. DoD map of the apparent difference for the pair S3-S1 (colours saturated at the minLoD = 0.488 m).

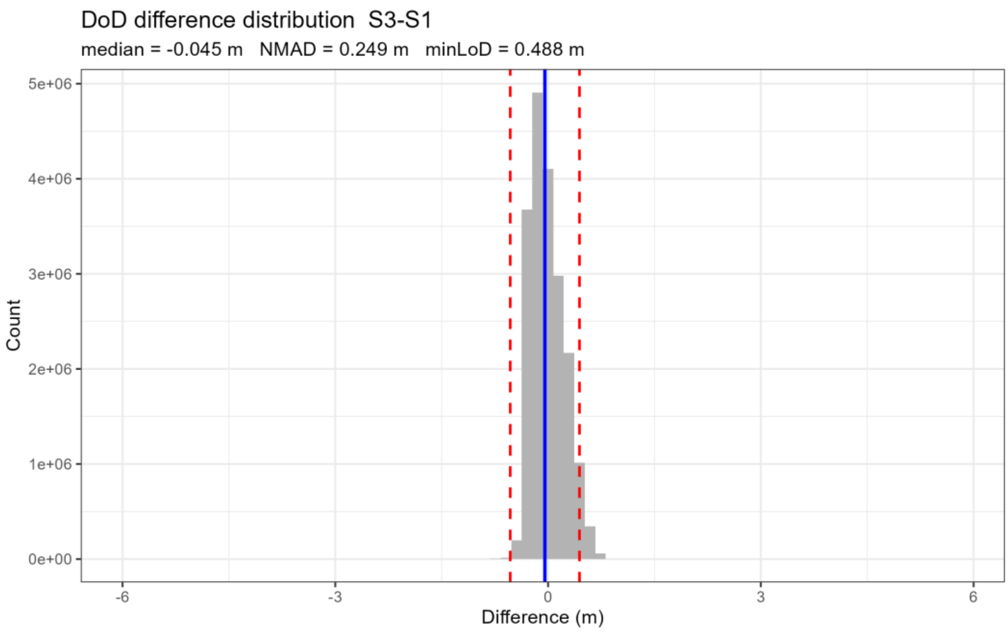

Figure 3.24. Distribution of the DoD differences for the pair S3-S1.

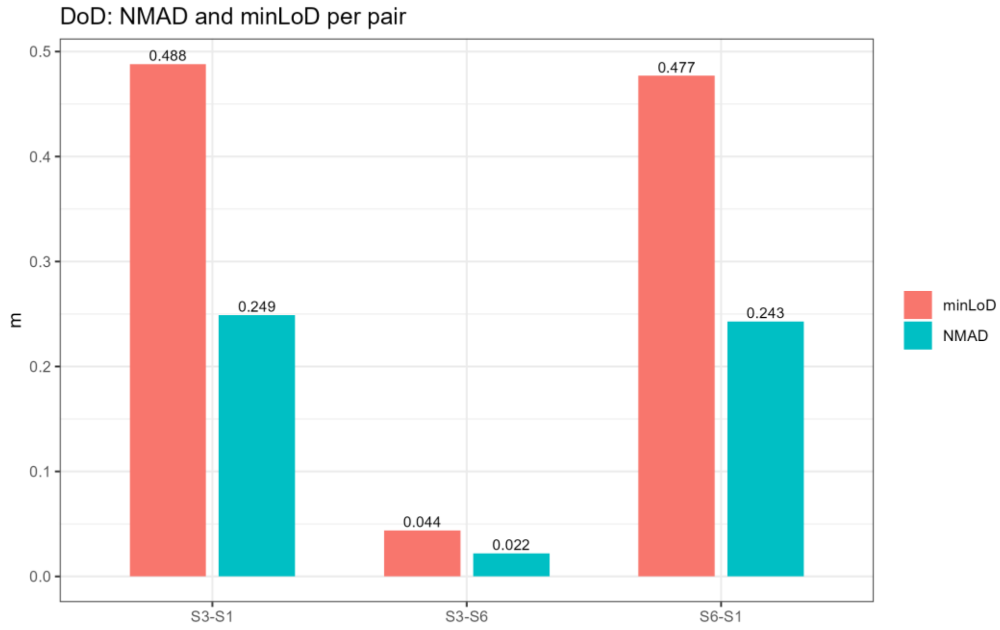

Figure 3.25. NMAD and minLoD of the DoD for each pair.

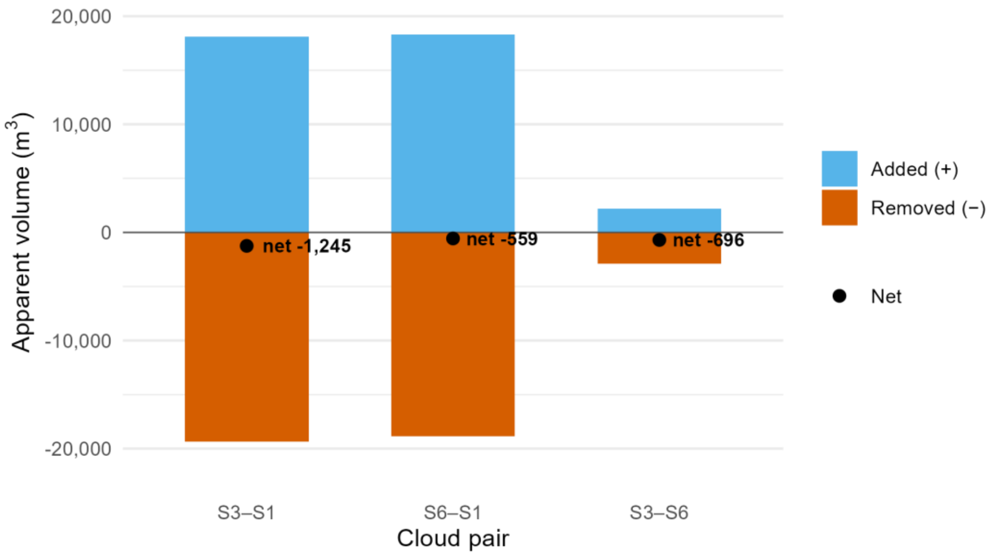

Figure 3.26. Apparent DoD volumes for each pair.

pair	Added	Removed	Net
S3-S1	18092.505	-19338.374	-1245.87
S6-S1	18311.175	-18870.433	-559.258
S3-S6	2194.407	-2890.987	-696.579

Table 3.12. Apparent DoD volumes (added, removed, net) for each pair.

4. Discussion

Models S1 and S3 cover the same terrain surveyed in the same acquisition campaign. Any difference between them is therefore not a real change of the surface, but the error introduced by the GPS georeferencing with respect to the Total Station. The null mean deviation shows that the low-cost model, as a whole, sits at the correct elevation: the correction through GCPs and the deactivation of the cameras have eliminated the overall vertical shift. A systematic tilt of 1.478° remains, however. The two models coincide at the centre but diverge progressively towards the edges, up to several metres of elevation at the extremities of the site. The cause is the drift of the GPS elevation during the survey, which translates into a rotation of the model. A rigid translation cannot straighten it, because it shifts but does not rotate. For this reason the tilt emerges only from the surface comparison and would not have been visible from the RMSE on the control points alone. That a single rotation explains 99.5% of the discrepancy is significant: the error of the workflow has a recognizable and coherent form, it is not random disorder.

Once the tilt is removed, the two surfaces agree to within about a quarter of a metre. This value sets the real noise floor of the method and, through the minLoD, its monitoring capability: the low-cost workflow reliably recognizes a change of the surface only if it exceeds about 0.5 m of vertical displacement. Smaller variations are confused with the error of the method itself. This leads to an important distinction for the research question. If one asks how accurate the low-cost is with respect to the gold standard, the answer includes the tilt, and the two models diverge by a few metres at the edges: GPS alone is not geometrically equivalent to the Total Station. If instead one asks whether the method serves multitemporal monitoring - the question of this thesis - then it is the residual that matters and not the tilt. Comparing two epochs surveyed with the same method, the systematic tilt tends to cancel, and what limits detectability is the threshold of about 0.5 m. In operational terms, the low-cost workflow is adequate to detect surface variations of the order of half a metre and above, not centimetric variations. The minLoD is a vertical-displacement threshold, expressed in linear metres: it indicates by how much the surface must have risen or lowered at a point for the change to be distinguishable from the noise. The estimation of erosion or deposition volumes, in cubic metres, derives instead from a subsequent computation that integrates the elevation differences over the cell areas.

4.1 GPS correction and photogrammetric accuracy

The characterization of the GPS error in § 3.1 shows that the residual between consumer coordinates and the Total Station reference is not a homogeneous random disturbance. It is the sum of three qualitatively distinct components. The planimetric component is dominated by a modest directional bias (about 0.6 m eastwards and 0.4 m northwards), with a random noise of the order of 0.35 m and no spatial structure ($r = 0.057$ between the planimetric residual and the distance to the nearest GCP). The elevation component is governed by a systematic offset of about 14 m, due to the difference between WGS84 ellipsoidal height and local orthometric height. To this is added a statistically robust temporal drift of ΔZ ($r = -0.713$), which makes the vertical bias non-constant during the measurement session. This decomposition directly conditions the choice of the correction model and explains its results.

The central result of § 3.2 follows from the geometry of the problem. The pure three-parameter translation (S3, mean 3D error of 0.862 m on the independent check points) outperformed the Helmert transformations (S4-S5, 0.911-0.962 m). Over an area of about 500 × 500 m, the rotation and scale components of Helmert are not separable from the translation. An infinitesimal rotation applied to UTM coordinates of the order of 4×10^6 m generates apparent translations as large as the real ones (§ 2.3.2; Li et al., 2012; Tran et al., 2023). The additional parameters therefore do not capture a real geometric structure; they fit the noise of the estimation sample. This is overfitting. Scenario S4 improves the fit on the GCPs only slightly, but worsens the accuracy on the independent check points, and in particular degrades CP 544 (from 0.582 m to 1.077 m). Empirically, this confirms the preference for the most parsimonious model when the control points do not allow the transformation components to be separated.

The ineffectiveness of the robust estimator with respect to the pure translation also deserves a specific reading. The global Huber estimator (S5) penalizes points with a high normalized residual. But here the dominant residual is the elevation one, which reflects the systematic temporal drift of the Z bias, not a measure of lower quality. Penalizing those points means discarding valid information, and indeed S5 does not improve on S4.

The planimetric residual error after correction (about 0.5 m, both on the GCPs and on the check points) is in line with what is documented for consumer GPS receivers in standalone mode (Kim et al., 2023). The dominant systematic component, the geoidal offset, is constant in space and therefore entirely removable with a single set of high-precision reference points. This is the very premise of the proposed workflow: to measure a reduced number of Total Station points only once and reuse them as a fixed reference. Alternative solutions based on differential GNSS or real-time kinematic (RTK)/post-processed kinematic (PPK) would give centimetric accuracy at the source (Štroner et al., 2021), but they are not feasible in the operational context of the region because of the regulatory and logistical constraints already discussed (Herrmann et al., 2018). The low-cost workflow evaluated here is therefore a pragmatic response to such constraints.

An unforeseen methodological outcome emerged in the photogrammetric processing and concerns the weight of the camera GPS positions in the bundle adjustment. When the images carry GPS coordinates with a strong elevation bias - here the elevations of the DJI Air 3S are systematically too low by about 100 m - the large number of camera observations (about 800) can prevail over the few control markers. It then drags the entire model towards a wrong elevation and tilts the surface systematically. The solution adopted, disabling the camera positions from the optimization and constraining the georeferencing to the markers alone (§ 2.5.1), reduces the model error from about a hundred metres to a few metres. It is an operational result, but transferable to any workflow that combines geotagged UAV images with control points, and it becomes a practical recommendation for low-cost surveys. The episode falls within a broader and well-documented problem: the large-scale systematic deformation of SfM topographic models. James and Robson (2014) show that DEMs from near-nadir UAV images undergo systematic deformations of a few decimetres over distances of about a hundred metres (p. 1419), all the more marked in SfM workflows with self-calibration and limited ground control (p. 1414). In their case the mechanism is the correlation, in parallel-acquisition networks, between the shape of the reconstructed surface and the estimation of the radial distortion, inseparable without additional information (p. 1416). Here the specific mechanism is different - the numerical dominance of the camera positions with elevation bias, not the radial distortion - but the principle is the same: the way the bundle adjustment treats the camera model and the georeferencing constraints can systematically deform the entire surface, with possible harm for comparative and multitemporal studies, and the nominal point accuracy alone does not reveal it. Hence the recommendation, reiterated by the same authors, to distribute the control points widely and to always use independent check points to detect and quantify these deformations (James & Robson, 2014, pp. 1419-1420).

The results finally require distinguishing two notions of accuracy that the analysis of the coordinates alone tends to confuse. The coordinate error of 0.862 m measures the deviation of the corrected positions at the check points. The accuracy of the reconstructed photogrammetric model - computed on the same check points after the bundle adjustment - is instead of the order of a few metres and lies entirely in the vertical component (§ 3.4.1). The gap does not indicate a worsening. It is the intended consequence of assigning the markers a loose elevation accuracy (Z = 20 m), which leaves the absolute elevation weakly constrained so as not to deform the model. The geometric fidelity that really matters for the analysis of surface variations, however, emerges from neither of the two point-wise metrics, but only from the direct comparison between surfaces, the subject of the following section.

4.2 Minimum number of GCPs and operational implications

The Monte Carlo analysis (§ 3.3) sets the operational threshold at seven GCPs, beyond which each additional control point gives a negligible benefit. With seven points, 96% of the planimetric correction attainable with all nineteen GCPs with respect to the uncorrected GPS is already recovered, and beyond the threshold the gain drops below 2% per additional point. The result separates two conceptually different benefits. The first is the reduction of the median RMSE, which is modest (from 0.538 m with three points to 0.491 m with nineteen). The second, more important operationally, is the narrowing of the 5-95% confidence band, which reduces by about 40% between n = 3 and n = 7. Adding points reduces above all the risk of a poor result due to an unfavourable selection of the sample, rather than the expected accuracy. This is consistent with the literature on GCP configuration in UAV photogrammetry (Sanz-Ablanedo et al., 2018; Villanueva & Blanco, 2019; Zhong et al., 2025), where the main gain of more points lies in reliability, not in mean accuracy.

Equally significant is the almost total invariance of the elevation RMSE with respect to the number of GCPs (range 0.896-0.924 m for each n; the lower bound of 0.896 m corresponds to an intermediate value of n not among the representative rows reported in Table 3.8). The residual vertical error is dominated by the temporal drift of the Z bias, a dynamic and non-random structure. For this reason, no increase in the number of static control points can reduce it. The correction would require explicitly modelling the drift over time, or a local geoidal reference, indicated among the development prospects (§ 5.3). The elevation limit is therefore one of modelling, not of sampling.

On the operational level, these outcomes matter directly for campaigns in resource-limited and access-constrained contexts such as the studied region (Herrmann et al., 2018). Measuring seven well-distributed points with a Total Station only once, and reusing them as a fixed reference in all subsequent campaigns, greatly reduces the cost and time of the precision geodetic survey without affecting the planimetric accuracy of the correction. The recommended distribution - perimeter points, at least one central point, and a coverage of the available elevation difference - follows the consolidated indications for UAV photogrammetry on variable morphology (Martínez-Carricondo et al., 2018; Türk et al., 2025, p. 85). Recent studies confirm this setup. Placing the points at the edges of the survey area reduces the error, while concentrating them at the centre worsens it, so that beyond a certain number the balance between quantity and distribution matters more than the count alone (Atik & Arkalı, 2025, pp. 13, 17). The elevation accuracy, in particular, improves with the number of control points and degrades as the distance from the nearest point increases, with a marked increase of the vertical error beyond 150 m (Deliry & Avdan, 2024, pp. 185-186). One firm point remains, however: the seven-point threshold is specific to the translation method and to the geometry of this site, and generalizing it to other contexts requires validation, as discussed among the limitations (§ 5.2).

4.3 Reliability of change detection

The assessment of the reliability of change detection rests on an argument that runs through the entire thesis: the accuracy of the coordinates at the control points is not enough to guarantee the geometric fidelity of the reconstructed surface. The bundle adjustment propagates the errors of the GCPs into the geometry of the model in ways that the point-wise residuals do not reveal (Liu et al., 2022; Zhong et al., 2025). In particular, a spatially structured elevation error - here the temporal drift of ΔZ - can systematically tilt the entire surface without the check on the check points noticing it. The surface comparison of § 3.4.2, conducted on the real dense clouds, made exactly this point clear: a systematic rotation of 1.478°, invisible at the coordinate level and identifiable only from the direct comparison between surfaces, which alone explains 99.5% of the raw discrepancy. This is in itself a methodological argument in favour of the surface-level assessment.

How to treat this systematic component depends on the analytical objective. To assess the georeferencing accuracy - the comparison between scenarios of the same acquisition campaign - the tilt is part of the result and must be reported, not removed, because it quantifies the geometric discrepancy of the low-cost workflow with respect to the gold standard. For the future multitemporal application the opposite holds: the systematic component shared by two epochs georeferenced on the same fixed points must be removed, so that the residual reflects the real change (§ 2.5.4). The two comparison methods adopted are complementary with respect to this problem. The DoD, computed along the vertical, is sensitive to slopes: a planimetric residual e, on a slope of gradient α, produces an apparent elevation difference of about $e \cdot \tan \alpha$. The M3C2, which measures along the normal to the surface, is instead little sensitive to this effect (Lague et al., 2013). Reporting both methods makes it possible to distinguish the real change from the artefacts due to the slope.

The minimum level of detection (minLoD) must be computed on the random component of the residual alone, after removing the systematic part, and with robust statistics - median and NMAD (Höhle & Höhle, 2009). The distribution of the residual is in fact not Gaussian, and the classical standard deviation would be inflated by the tilt and the outliers. The operational answer to the research question depends on the comparison between this threshold and the erosion expected on the flanks of the tell. If the minLoD is lower than the surface variation expected between two campaigns, the low-cost workflow is suitable for multitemporal analysis; otherwise the uncertainty of the survey masks the signal. This comparison is now quantified on the definitive models (§ 3.4.2, Table 3.9). Once the systematic tilt is removed, the random residual between the surfaces is of decimetric scale (NMAD ≈ 0.22-0.27 m), from which a minLoD of the order of 0.5 m of vertical displacement. The low-cost workflow therefore reliably distinguishes only the surface changes that exceed this threshold, while the smaller variations remain confused with the noise of the method. The detailed interpretation - and in particular why the reduction of the GCPs does not degrade multitemporal detectability - is taken up in § 4.4.

4.4 Accuracy, bias, and detection capability

The results of the three pairs require a clarification, because two distinct indicators risk being confused. The minLoD does not measure the accuracy of a scenario with respect to the Total Station reference, but only the random noise that remains after the removal of the systematic components. The accuracy with respect to the gold standard is instead described by the mean bias and the tilt. Keeping the two levels separate avoids a paradoxical reading of Table 3.9, in which S6 would seem to have a better detection threshold than S3.

Where S3 is actually more accurate is in the mean bias. With nineteen control points the scenario is almost perfectly centred on the reference (0.168 m), while with seven points a systematic elevation deviation of about half a metre remains (−0.538 m). This is the expected effect of the reduction of the control points: a larger number of points constrains the correction better and reduces the systematic error. The advantage of the nineteen points manifests itself in this component, not in the random noise.

The residual random noise, instead, is substantially identical in the two scenarios (about 0.24 m) and does not depend on the number of control points. This noise floor is set by the photogrammetric reconstruction and by the planimetric residual projected onto the sloping flanks, not by the alignment to the control points: the points correct the systematic error, not the random dispersion. That the residual does not drop to the centimetric scale despite the high density of the clouds indicates that it is not matching noise, but a real geometric limit of the workflow, which denser clouds would not reduce. The lower apparent threshold of S6 in Table 3.9 derives solely from the use of the NMAD instead of the standard deviation; computed on σ, the threshold of the two scenarios is in fact identical (about 0.47-0.48 m).

From this follows the conclusion relevant to the research question. Reducing the control points from nineteen to seven worsens the mean bias but does not degrade the detection capability in multitemporal monitoring, because bias and tilt cancel between two epochs surveyed with the same method. The S3-S6 pair shows this directly: comparing two corrections of the same type, the common tilt cancels and the dispersion drops to 0.133 m. The capacity to distinguish a real change therefore depends on the random residual of the workflow, common to both configurations, and not on the number of control points employed. This reading finds support in the literature on deformation monitoring: even a high density of control points acts above all as a safety margin, while with well-chosen processing parameters a high accuracy is obtained with a much smaller number of points, and - a central aspect for the multitemporal case - the difference between two epochs suppresses the low-frequency systematic component, making the optimization of the parameters less critical for the derived quantities (Ćwiąkała et al., 2026, pp. 12, 14, 15).

5. Conclusions

5.1 Summary of the results

The thesis evaluated to what extent a low-cost georeferencing workflow - consumer GPS corrected with a limited number of Total Station points - can produce UAV models suitable for analysing multitemporal surface variations at an archaeological site under real operational conditions. The characterization of the GPS error revealed a three-component structure. First, a modest planimetric bias with no spatial structure. Second, a systematic geoidal offset of about 14 m. Third, a temporal drift of the elevation bias. Among the correction models evaluated, the pure three-parameter translation proved optimal, with a mean 3D error of 0.862 m on the independent check points. It outperformed the Helmert transformations, which were penalized by the non-separability of the parameters over a small area. The Monte Carlo analysis identified seven Total Station points as the operational threshold that recovers 96% of the planimetric correction with respect to the uncorrected GPS. The gain lies above all in reliability rather than in mean accuracy. The surface-level accuracy assessment, conducted with the M3C2 algorithm on the dense clouds of scenarios S1, S3, and S6 (§ 3.4.2), provided the quantitative answer to the research question. The comparison reveals a common systematic component: a tilt of about 1.478° produced by the drift of the GPS elevation, which alone explains 99.5% of the discrepancy. Once this tilt is removed, a random residual of about 0.24 m remains. From this follows a minimum level of detection (minLoD) of the order of 0.5 m of vertical displacement. The low-cost workflow is therefore suitable for detecting surface variations of half a metre and above, not centimetric variations. The systematic tilt and the mean bias cancel between epochs surveyed with the same method, as the internal consistency comparison S3-S6 shows (§ 3.4.2). The multitemporal detection capability is therefore determined by the random residual alone, not by the number of control points employed.

5.2 Limitations and methodological contributions

The work presents some methodological limitations that circumscribe its scope. The Monte Carlo analysis verifies only the translation method (S3), already identified as optimal, not the entire set of correction models; the seven-GCP threshold is moreover specific to the geometry of this site. The correction of the non-common points uses the traditional model, which ignores the uncertainty propagated by the transformation estimated on the common GCPs. The rigorous seamless model of Li et al. (2012) would give a more complete estimate, but was not implemented. The temporal drift of the elevation bias, attributable to ionospheric effects on the single-frequency L1 signal (Štroner et al., 2021), was neither modelled nor corrected. The photogrammetric processing was conducted at Medium dense-cloud quality - a choice consistent with the constraints of the low-cost workflow, but sub-optimal in absolute terms. It produced a vertical residual of a few metres at the model level, due to the deliberately loose elevation constraint. Finally, the work is based on a single site and a single epoch. The real multitemporal comparison, which requires the North-South flight block not yet processed or subsequent campaigns, remains to be carried out. A limitation of the comparison method itself must also be considered. On rough surfaces such as the flanks of the tell, the standard M3C2 tends to overestimate the local uncertainty, and therefore the detection threshold, while the subsampling of the core points (here 0.25 m) reduces its spatial resolution. More recent variants based on planar patches (PB-M3C2) attenuate both effects and indicate a direction of refinement for multitemporal monitoring (Yang & Schwieger, 2023, pp. 3, 8).

On the level of the volumetric analysis, the present work reports the overall apparent volume of the DoD without applying the minimum level of detection to the individual cells. The framework of Wheaton et al. (2010, p. 140) provides for a thresholded DoD, in which the cells whose difference does not exceed the minLoD are excluded from the computation, so as to separate the real change volume from the noise. Such thresholding was not applied here because the three comparisons concern the same epoch. The measured volume is entirely apparent - the artefact of the georeferencing difference alone - and its thresholding would not correspond to any physical change. The application of the thresholded DoD, and with it the propagation of the threshold between two epochs, is therefore deferred to the real multitemporal comparison, where it would acquire meaning.

On the side of the contributions, the thesis proposes and documents a reproducible operational workflow for the low-cost georeferencing of UAV surveys in access-constrained contexts. The workflow includes two practical measures that emerged from the analysis. The first is disabling the camera positions from the optimization when the GPS elevations have a systematic bias. The second is assigning the markers an asymmetric XY/Z accuracy to avoid the deformation of the model. On the methodological level, the thesis also shows that the accuracy of the coordinates alone is not enough to characterize the geometric fidelity of the models. The surface-level comparison serves to reveal systematic components - such as the tilt induced by the elevation drift - otherwise invisible.

5.3 Prospects for future research

Several directions of development emerge from the work. The most immediate is to build a local geoidal model to systematically correct the elevation bias. This would address at the root the vertical limitation, which increasing the number of GCPs does not reduce. Integrating the North-South flight block and subsequent campaigns would make it possible to move from the accuracy analysis to real multitemporal change detection, with the estimation of the volumetric budgets of the tell's erosion. On the side of the correction methods, extending the Monte Carlo analysis to all the models would quantify the behaviour of the robust variants as the sample varies. For the surface comparison, adopting the explicit point-by-point uncertainty propagation in M3C2 (M3C2-EP; Winiwarter et al., 2021) would make the error budget fully explicit. Developing a workflow for correcting the systematic tilt between epochs would complete the multitemporal pipeline.

A further prospect concerns the rigorous treatment of the non-common points - the GPS points to which the transformation estimated on the common GCPs is applied. In scenarios S3-S5 this transformation, estimated on the common GCPs, is then applied to the non-common points without taking into account the additional uncertainty that the propagation introduces. Li et al. (2012, p. 2107) show that the traditional model is a reduced case of the rigorous seamless model and, as such, ignores the errors of the coordinates of the non-common points during the transformation. The seamless model corrects them by taking into account their spatial correlation with the common points. Where such spatial correlation is significant, the transformed coordinates of the non-common points would improve with respect to the traditional model. Li et al. (2012, p. 2107) specify, however, that the two models give comparable results when the common points have high precision and weak correlations with the non-common points. Here the limited precision of the consumer GPS ($\sigma \approx 0.35$ m) indicates that the conditions for comparability might not be fully satisfied. Implementing the seamless model of Li et al. (2012) therefore represents a priority methodological improvement for future developments: it would give a more rigorous estimate of the uncertainty of the corrected coordinates of the non-common points.

References

Atik, M. E., & Arkalı, M. (2025). Comparative assessment of the effect of positioning techniques and ground control point distribution models on the accuracy of UAV-based photogrammetric production. Drones, 9(1), 15. https://doi.org/10.3390/drones9010015

Besl, P. J., & McKay, N. D. (1992). A method for registration of 3-D shapes. IEEE Transactions on Pattern Analysis and Machine Intelligence, 14(2), 239-256. https://doi.org/10.1109/34.121791

Campana, S. (2017). Drones in archaeology. State-of-the-art and future perspectives. Archaeological Prospection, 24(4), 275-296. https://doi.org/10.1002/arp.1569

Cunliffe, E. L., de Gruchy, M. W., & Stammitti, E. (2012). How to build a dam and save cultural heritage. International Journal of Heritage in the Digital Era, 1(1_suppl), 221-226. https://doi.org/10.1260/2047-4970.1.0.221

Ćwiąkała, P., Puniach, E., Pastucha, E., & Gruszczyński, W. (2026). The impact of processing parameters on high-accuracy measurements in UAV photogrammetry. Measurement, 265, 120315. https://doi.org/10.1016/j.measurement.2026.120315

Deliry, S. I., & Avdan, U. (2024). Accuracy assessment of UAS photogrammetry and structure from motion in surveying and mapping. International Journal of Engineering and Geosciences, 9(2), 165-190. https://doi.org/10.26833/ijeg.1366146

Eggert, D. W., Lorusso, A., & Fisher, R. B. (1997). Estimating 3-D rigid body transformations: A comparison of four major algorithms. Machine Vision and Applications, 9(5-6), 272-290. https://doi.org/10.1007/s001380050048

Fiz, J. I., Martín, P. M., Cuesta, R., Subías, E., Codina, D., & Cartes, A. (2022). Examples and results of aerial photogrammetry in archeology with UAV. Drones, 6(3), 59. https://doi.org/10.3390/drones6030059

Forti, L., Perego, A., & Brandolini, F. (2021). Geomorphology of the northwestern Kurdistan Region of Iraq. Journal of Maps, 17(2), 225-236. https://doi.org/10.1080/17445647.2021.1906339

Forti, L., Mariani, G. S., Brandolini, F., Pezzotta, A., & Zerboni, A. (2022). Declassified intelligence satellite imagery as a tool to reconstruct past landforms: The submerged riverscape of the Tigris River below the Mosul Dam Lake. Earth Surface Processes and Landforms, 47(10), 2483-2499. https://doi.org/10.1002/esp.5389

Forti, L., Costanzo, S., Compostella, C., Garna, G., Morandi Bonacossi, D., & Zerboni, A. (2023a). The geoarchaeological investigation on the defunctionalisation of an Assyrian canals system reveals Late-Holocene land use transitions in Northern Mesopotamia. The Holocene, 33(4), 416-431. https://doi.org/10.1177/09596836221145395

Forti, L., Brandolini, F., Oselini, V., Peyronel, L., Pezzotta, A., Vacca, A., & Zerboni, A. (2023b). Geomorphological assessment of the preservation of archaeological tell sites. Scientific Reports, 13, 7683. https://doi.org/10.1038/s41598-023-34490-4

Forti, L., Pezzotta, A., Zebari, M., & Zerboni, A. (2023c). Geomorphology of the central Kurdistan Region of Iraq. Journal of Maps, 19(1). https://doi.org/10.1080/17445647.2022.2157306

Google. (n.d.). [Satellite imagery of the Mosul Dam reservoir, Iraq] [Map]. Retrieved 08/08/2026, from Google Satellite (loaded via the HCMGIS plugin in QGIS).

Gräler, B., Pebesma, E., & Heuvelink, G. (2016). Spatio-temporal interpolation using gstat. The R Journal, 8(1), 204-218. https://doi.org/10.32614/RJ-2016-014

He, Y., Lei, S., Dai, W., Chen, X., Wang, B., Sheng, Y., & Lin, H. (2025). DEM-based topographic change detection considering the spatial distribution of errors. Geo-spatial Information Science, 28(2), 497-510. https://doi.org/10.1080/10095020.2024.2324921

Herrmann, J. T., Glissmann, B., Sconzo, P., & Pfälzner, P. (2018). UAV survey with commercial-grade instruments: A case study from the Eastern Ḫabur Archaeological

Survey, Iraq. Journal of Field Archaeology, 43(4), 269-283. https://doi.org/10.1080/00934690.2018.1465808

Höhle, J., & Höhle, M. (2009). Accuracy assessment of digital elevation models by means of robust statistical methods. ISPRS Journal of Photogrammetry and Remote Sensing, 64(4), 398-406. https://doi.org/10.1016/j.isprsjprs.2009.02.003

Huber, P. J. (1992). Robust estimation of a location parameter. In S. Kotz & N. L. Johnson (Eds.), Breakthroughs in statistics (pp. 492-518). Springer. https://doi.org/10.1007/978-1-4612-4380-9_35

James, M. R., & Robson, S. (2014). Mitigating systematic error in topographic models derived from UAV and ground-based image networks. Earth Surface Processes and Landforms, 39(10), 1413-1420. https://doi.org/10.1002/esp.3609

Kim, H., Hyun, C.-U., Park, H.-D., & Cha, J. (2023). Image mapping accuracy evaluation using UAV with standalone, differential (RTK), and PPP GNSS positioning techniques. Sensors, 23(13), 5858. https://doi.org/10.3390/s23135858

Lague, D., Brodu, N., & Leroux, J. (2013). Accurate 3D comparison of complex topography with terrestrial laser scanner. ISPRS Journal of Photogrammetry and Remote Sensing, 82, 10-26. https://doi.org/10.1016/j.isprsjprs.2013.04.009

Li, B., Shen, Y., & Li, W. (2012). The seamless model for three-dimensional datum transformation. Science China Earth Sciences, 55(12), 2099-2108. https://doi.org/10.1007/s11430-012-4418-z

Liu, X., Lian, X., Yang, W., Wang, F., Han, Y., & Zhang, Y. (2022). Accuracy assessment of a UAV direct georeferencing method and impact of the configuration of ground control points. Drones, 6(2), 30. https://doi.org/10.3390/drones6020030

Lucks, L., Stilla, U., Hoegner, L., & Holst, C. (2024). Photogrammetric rockfall monitoring in Alpine environments using M3C2 and tracked motion vectors. ISPRS Open Journal of Photogrammetry and Remote Sensing, 12, 100058. https://doi.org/10.1016/j.ophoto.2024.100058

Marchetti, N., Curci, A., Gatto, M. C., Nicolini, S., Mühl, S., & Zaina, F. (2019). A multi-scalar approach for assessing the impact of dams on the cultural heritage in the Middle East and North Africa. Journal of Cultural Heritage, 37, 17-28. https://doi.org/10.1016/j.culher.2018.10.007

Martínez-Carricondo, P., Agüera-Vega, F., Carvajal-Ramírez, F., Mesas-Carrascosa, F.-J., García-Ferrer, A., & Pérez-Porras, F.-J. (2018). Assessment of UAV-photogrammetric mapping accuracy based on variation of ground control points. International Journal of Applied Earth Observation and Geoinformation, 72, 1-10. https://doi.org/10.1016/j.jag.2018.05.015

Mercan, H., Akyilmaz, O., & Aydin, C. (2018). Solution of the weighted symmetric similarity transformations based on quaternions. Journal of Geodesy, 92(10), 1113-1130. https://doi.org/10.1007/s00190-017-1104-0

Over, J.-S. R., Ritchie, A. C., Kranenburg, C. J., Brown, J. A., Buscombe, D. D., Noble, T., Sherwood, C. R., Warrick, J. A., & Wernette, P. A. (2021). Processing coastal imagery with Agisoft Metashape Professional Edition, version 1.6. U.S. Geological Survey Open-File Report 2021-1039. https://doi.org/10.3133/ofr20211039

Puniach, E., Ćwiąkała, P., Wasilewski, M., Majchrzak, Ł., Guillen, P., Escobar, C., & Marchewka-Długońska, J. (2026). Monitoring archaeological sites affected by natural and anthropogenic hazards using multitemporal UAV-photogrammetry: A case study from the Barranca Valleys, Peru. Journal of Field Archaeology, 51(3), 159-178. https://doi.org/10.1080/00934690.2026.2619809

R Core Team. (2024). R: A language and environment for statistical computing. R Foundation for Statistical Computing. https://www.R-project.org/

Rosen, A. M. (1986). Cities of Clay: The Geoarchaeology of Tells. University of Chicago Press.

Sanz-Ablanedo, E., Chandler, J. H., Rodríguez-Pérez, J. R., & Ordóñez, C. (2018). Accuracy of unmanned aerial vehicle (UAV) and SfM photogrammetry survey as a function of the number and location of ground control points. Remote Sensing, 10(10), 1606. https://doi.org/10.3390/rs10101606

Sconzo, P. (2024). Alla riscoperta dei paesaggi sommersi del bacino della diga di Mosul: Il Progetto archeologico ReLand. Mare Internum, 16, 145-154. https://doi.org/10.19272/20243140100

Sconzo, P., & Simi, F. (2020). Settlement dynamics on the banks of the upper Tigris, Iraq. Journal of Open Archaeology Data, 8(3). https://doi.org/10.5334/joad.632019

Sconzo, P., & Simi, F. (2024). Waiting for the deluge? Some thoughts on the impact of dams on the heritage in Southwest Asia and North Africa. Mesopotamia, LIX, 85-108.

Sconzo, P., Simi, F., & Titolo, A. (2023). Drowned landscapes: The rediscovered archaeological heritage of the Mosul Dam Reservoir. Bulletin of the American Society of Overseas Research, 389, 165-189. https://doi.org/10.1086/724419

Shapiro, S. S., & Wilk, M. B. (1965). An analysis of variance test for normality (complete samples). Biometrika, 52(3-4), 591-611. https://doi.org/10.1093/biomet/52.3-4.591

Štroner, M., Urban, R., Seidl, J., Reindl, T., & Brouček, J. (2021). Photogrammetry using UAV-mounted GNSS RTK: Georeferencing strategies without GCPs. Remote Sensing, 13(7), 1336. https://doi.org/10.3390/rs13071336

Tran, D. T., Nocquet, J.-M., Luong, N. D., & Nguyen, D. H. (2023). Determination of Helmert transformation parameters for continuous GNSS networks. Geo-spatial Information Science, 26(1), 125-138. https://doi.org/10.1080/10095020.2022.2138569

Türk, T., Bahadur, B., Demirel, Y., Altuntaş, C., & Öcalan, T. (2025). Investigation of the effect of GCP number and distribution on photogrammetric product accuracy in UAV photogrammetry. Journal of Geodesy and Geoinformation, 12(1), 77-88. https://doi.org/10.9733/JGG.2025R0006.E

Villanueva, J. K. S., & Blanco, A. C. (2019). Optimization of ground control point (GCP) configuration for UAV survey using SfM. ISPRS Archives, XLII-4/W12, 117-124. https://doi.org/10.5194/isprs-archives-XLII-4-W12-117-2019

Wheaton, J. M., Brasington, J., Darby, S. E., & Sear, D. A. (2010). Accounting for uncertainty in DEMs from repeat topographic surveys. Earth Surface Processes and Landforms, 35(2), 136-156. https://doi.org/10.1002/esp.1886

Wilkinson, T. J. (2003). Archaeological Landscapes of the Near East. University of Arizona Press.

Wilson, G., Bryan, J., Cranston, K., Kitzes, J., Nederbragt, L., & Teal, T. K. (2017). Good enough practices in scientific computing. PLOS Computational Biology, 13(6), e1005510. https://doi.org/10.1371/journal.pcbi.1005510

Winiwarter, L., Anders, K., & Höfle, B. (2021). M3C2-EP: Pushing the limits of 3D topographic point cloud change detection by error propagation. ISPRS Journal of Photogrammetry and Remote Sensing, 178, 240-258. https://doi.org/10.1016/j.isprsjprs.2021.06.011

Yang, Y., & Schwieger, V. (2023). Patch-based M3C2: Towards lower-uncertainty and higher-resolution deformation analysis of 3D point clouds. International Journal of Applied Earth Observation and Geoinformation, 125, 103535. https://doi.org/10.1016/j.jag.2023.103535

Zhong, H., Duan, Y., Tao, P., & Zhang, Z. (2025). Influence of ground control point reliability and distribution on UAV photogrammetric 3D mapping accuracy. Geo-spatial Information Science, 28(5), 1998-2018. https://doi.org/10.1080/10095020.2025.2451204

Appendix A - Tables (detailed parameters and results)

parameter	value_m
tE	-0.5962
tN	-0.451
tZ	14.0831

Table A1. S3 translation parameters

parameter	value
scale	0.999457
dev_ppm	-543.243109
rot_X_deg	0.083186
rot_Y_deg	0.17053
rot_Z_deg	0.064968
tE_m	-0.553483
tN_m	-0.471852
tZ_m	14.076538

Table A2. S4 Helmert parameters

parameter	value
scale	0.999814
rot_X_deg	0.099424
rot_Y_deg	0.142163
rot_Z_deg	0.080933
tE_m	-0.585199
tN_m	-0.502091
tZ_m	14.152497
sigma_MAD_m	0.401735

Table A3. Robust Helmert S5 parameters

point	weight	residual_m
100	0.3639	1.4848
300	0.4671	1.1568
200	0.487	1.1095
400	0.5331	1.0137
527	0.6701	0.8064
531	1	0.5108
530	0.7591	0.7118
545	1	0.3547
546	0.9242	0.5846
543	0.4197	1.2873000000000001
542	1	0.4329

point	weight	residual_m
541	1	0.4474
540	0.4155	1.3004
539	1	0.4314
538	0.5208	1.0376
537	0.7693	0.7024
536	1	0.4616
535	1	0.4467
534	1	0.242

Table A4. Huber weights per point

scenario	check_point	res_E	res_N	res_Z	res_3D
S2 GPS raw	526	0.7171	-0.1948	-12.5768	12.5988
S2 GPS raw	533	0.9553	0.4268	-14.0669	14.1057
S2 GPS raw	544	0.2311	0.6241	-14.5017	14.517
S3 Translation	526	0.1208	-0.6458	1.5063	1.6434
S3 Translation	533	0.3591	-0.0242	0.0163	0.3603
S3 Translation	544	-0.3651	0.1731	-0.4186	0.5818
S4 Helmert	526	0.1057	-0.6071	1.3668	1.4993
S4 Helmert	533	0.0539	-0.0718	0.1285	0.1567
S4 Helmert	544	-0.2048	0.3818	-0.9861	1.077
S5 Robust Helmert	526	0.0917	-0.6171	1.4698	1.5967
S5 Robust Helmert	533	-0.0118	0.0097	0.3094	0.3098
S5 Robust Helmert	544	-0.1442	0.3184	-0.9149	0.9794

Table A5. Check point residuals per scenario

parameter	value_m
tE	-0.6556
tN	-0.4803
tZ	14.3383

Table A6. S6 translation parameters (7 GCP)

parameter	S3_19GCP_m	S6_7GCP_m	diff_m
tE	-0.5962	-0.6556	-0.0593
tN	-0.451	-0.4803	-0.0293
tZ	14.0831	14.3383	0.2551

Table A7. Comparison of S3 vs S6 parameters

check_point	res_E	res_N	res_Z	res_3D
526	0.0615	-0.6751	1.7614	1.8874
533	0.2998	-0.0535	0.2714	0.4079
544	-0.4244	0.1438	-0.1635	0.477

Table A8. S6 check point residuals

check_point	S3_3D_m	S6_3D_m	diff_m
526	1.6434	1.8874	0.244
533	0.3603	0.4079	0.0476
544	0.5818	0.477	-0.1048
Mean	0.8618	0.9241	0.0623

Table A9. Comparison of S3 vs S6 residuals

Site	Area (ha)	Site	Area (ha)
ReL_07	18.99	ReL_60	1.16
ReL_23	18.86	ReL_27	1.11
ReL_43	18.28	ReL_69	1.10
ReL_57	15.01	ReL_24	1.10
ReL_74	13.58	ReL_54	1.06
ReL_42	12.96	ReL_30	0.96
ReL_41	12.18	ReL_83	0.92
ReL_12	12.13	ReL_25	0.86
ReL_59	10.46	ReL_13	0.78
ReL_05	8.33	ReL_15	0.78
ReL_45	7.58	ReL_71	0.69
ReL_51	5.75	ReL_03	0.48
ReL_34	5.67	ReL_78	0.48
ReL_26	5.06	ReL_29	0.47
ReL_09	4.83	ReL_76	0.44
ReL_17	4.45	ReL_64	0.43
ReL_40	4.29	ReL_77	0.42
ReL_37	4.05	ReL_82	0.41
ReL_06	3.99	ReL_63	0.40
ReL_28	3.67	ReL_32	0.39
ReL_81	3.54	ReL_10	0.36
ReL_39	3.53	ReL_85	0.24
ReL_08	3.42	ReL_33	0.22
ReL_67	3.31	ReL_66	0.21
ReL_04	3.25	ReL_38	0.20
ReL_16	3.13	ReL_72	0.19
ReL_53	3.13	ReL_50	0.17
ReL_02	3.02	ReL_84	0.17
ReL_01	2.90	ReL_49	0.16
ReL_21	2.87	ReL_48	0.14
ReL_44	2.58	ReL_47	0.11
ReL_68	2.49	ReL_62	0.08
ReL_18	2.18	ReL_80	0.06

Site	Area (ha)	Site	Area (ha)
ReL_36	1.73	ReL_56	0.05
ReL_46	1.71	ReL_79	0.04
ReL_14	1.67	ReL_52	0.02
ReL_70	1.67	ReL_35	0.01
ReL_11	1.57	ReL_61	0.01
ReL_55	1.53	ReL_86	0.01
ReL_31	1.43	ReL_22	0.01
ReL_65	1.25	ReL_73	0.01
ReL_20	1.21	ReL_19	0.00
ReL_58	1.17	ReL_75	0.00

Table A10. All sites surveyed by the ReLand mission in the Mosul Dam basin (n = 86), with areas recomputed from the site polygon coordinates (EPSG:32638), ranked by area.

Appendix B - Figures

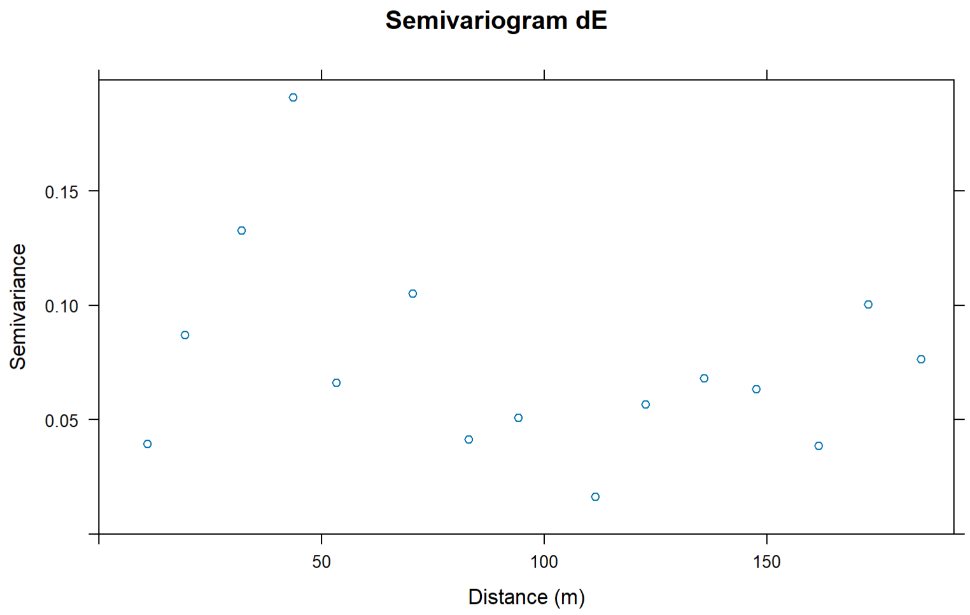

Figure B1. dE semivariogram

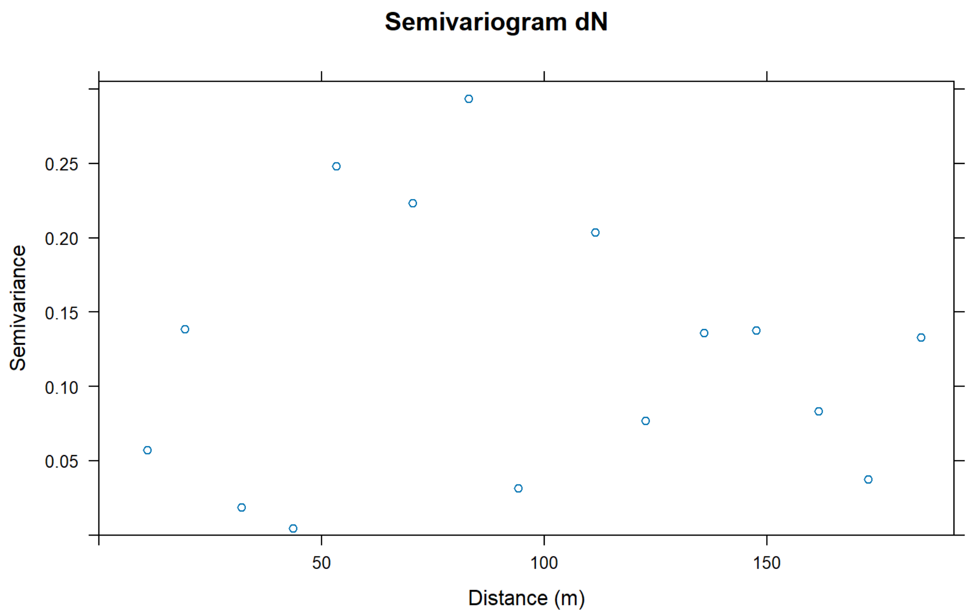

Figure B2. dN semivariogram

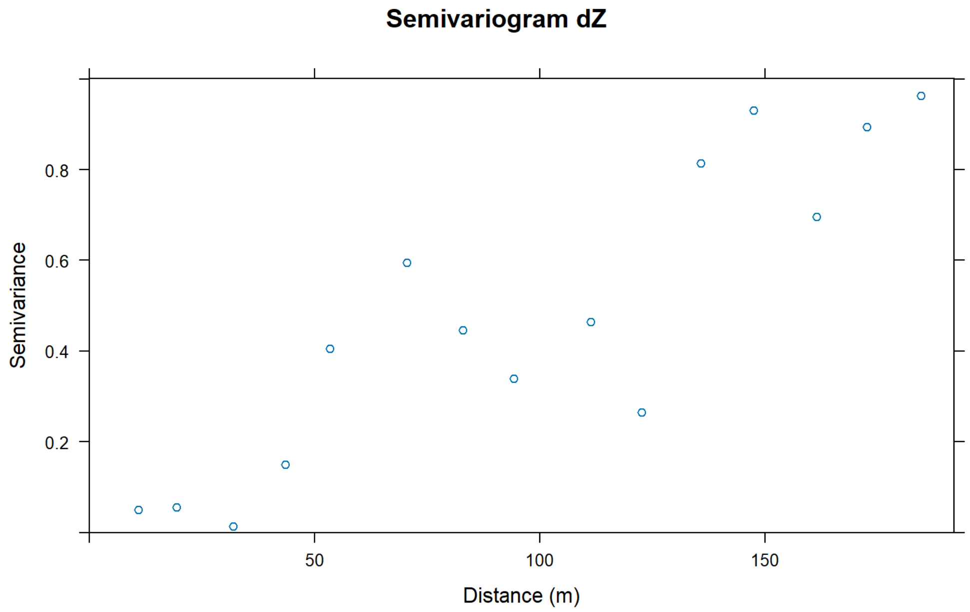

Figure B3. dZ semivariogram

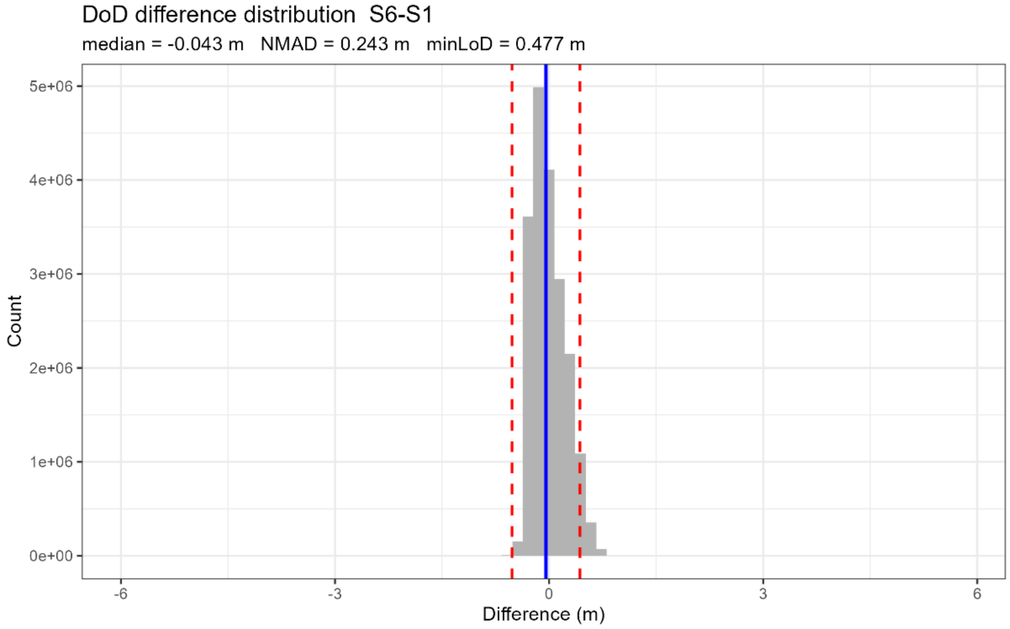

Figure B4. DoD difference distribution S6-S1 (replicate)

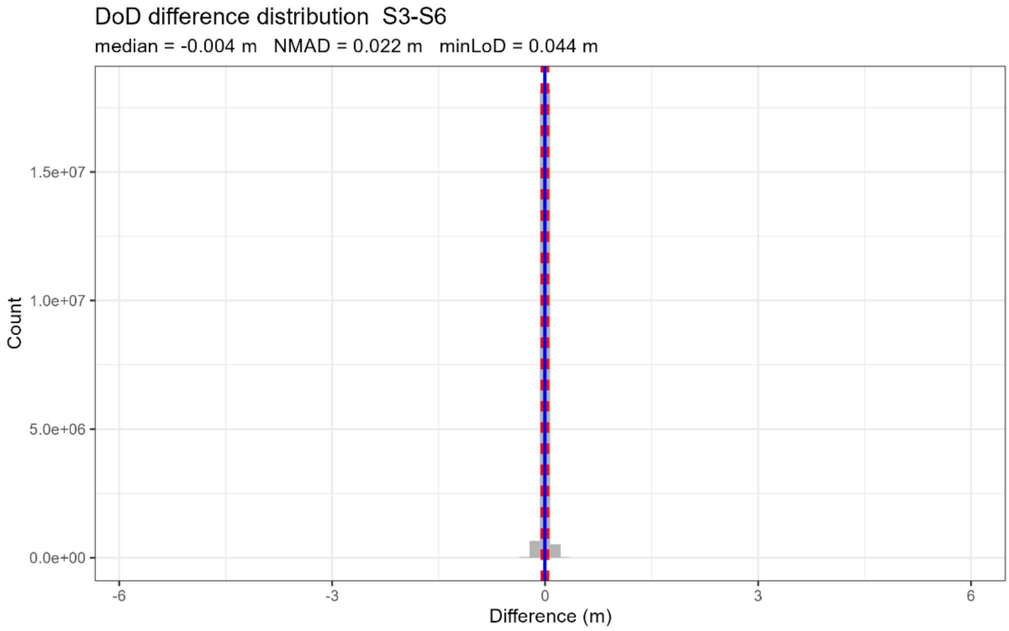

Figure B5. DoD difference distribution S3-S6 (replicate)