Article title

Automated grain sizing from UAV imagery of a gravel-bed river: benchmarking of three object-based methods and spatial analysis of grain size distributions

Abstract

Measuring grain sizes in gravel-bed rivers is crucial when studying river dynamics and sediment transport. Automated methodologies have been developed in recent years for detecting individual grains and measuring their size on digital imagery. These object-based methodologies have mainly been applied to handheld imagery. Low-cost and high-resolution orthoimages covering long river reaches are nowadays accessible with the improvements in Uncrewed Aerial Vehicles (UAV) and Structure-from-Motion (SfM) photogrammetry. Applying object-based grain sizing methodologies to such orthoimages may provide wide-scale information about the grain-size spatial distribution along streambeds. We first examined how accurate three object-based models (BASEGRAIN, PebbleCountsAuto and GALET) were, by comparing their outcomes to in-field manual measurements of grain sizes and manual grain labelling. We found that BASEGRAIN and PebbleCountsAuto underestimated grain sizes on average, whereas GALET generally overestimated grain size percentiles. Grain size measurements obtained by manually labelling grain features were consistent with in-field measurements. We then show that spatial statistics applied to automatically detected grain features allowed us to draw information about the grain-size organisation in an Alpine braided river. Spatial statistics were instrumental in consistently identifying patches of different grain sizes and thereby provided evidence for marked grain-size patchiness.

Keywords

Automated grain-size measurements; UAV; spatial statistics; grain-size organisation; gravel-bed rivers.

1. INTRODUCTION

River beds are seldom composed of sediments of uniform size. Quantifying the size distribution of sediment mixtures is fundamental for understanding and modelling river flows, morphodynamics and sediment transport processes (Bunte and Abt, 2001). For instance, a number of flow resistance equations involve grain size quantiles (e.g., Strickler, 1923; Keulegan, 1938; Hey, 1979; Smart and Jaeggi, 1983; Ferguson, 2007; Rickenmann and Recking, 2011). Another example is provided by the methods for predicting the threshold of incipient motion and transport rates (e.g., Recking, 2013; Recking, 2016; Hodge et al., 2007). When studying bedload transport in mountain rivers—whose sediment deposits are often prone to grain sorting processes, such as bed armouring and size-dependent selective transport—accurate quantification of local grain size distributions is particularly important for assessing bedload transport mechanisms (Schlunegger et al., 2020).

A variety of field sampling procedures have been developed for characterising river-bed grain-sizes in quantitative terms. Mechanical sieving of volumetric samples and individual measurements of surface grains are commonly used techniques (Kellerhals and Bray, 1971). Nonetheless, these procedures can be laborious and entail limitations for the assessment of spatial variations of bed-material grain-sizes (Bunte and Abt, 2001). To mitigate these issues, automated grain sizing methods based on remote sensing technologies have emerged in recent decades. A number of research studies have been dedicated to utilizing images for measuring grain sizes (see Carrivick and Smith, 2019; Piégay et al., 2020). Parallelly, recent improvements in Uncrewed Aerial Vehicles (UAV) and Structure-from-Motion (SfM) photogrammetry software packages have allowed to produce easily and at relatively low-cost valuable topographic datasets, which typically comprise orthoimages, digital surface models (DSM) and dense point clouds. Combining UAV-SfM imagery and image-based grain sizing techniques can provide access to grain size estimates over large spatial scales, in a cost- and time-efficient manner (Carrivick and Smith, 2019).

The first image-based methodologies for measuring grain sizes in rivers relied on the visual interpretation of nadir photographs (e.g., Adams, 1979; Ibbeken and Schleyer, 1986). Grain size distributions (GSD) were estimated by measuring the projected intermediate axis (b-axis) of grains on photographs. Although less field time is required (only photographs need to be taken), this method still necessitates large processing times to measure grain features individually (Bunte and Abt, 2001). Since the early 2000s, numerous studies were dedicated to automatizing grain sizing procedures on images (McEwan et al., 2000; Butler et al., 2001; Graham et al., 2005). Two distinct approaches have been followed. The first approach involves deriving grain sizes from image statistics (Carbonneau et al., 2004; Buscombe, 2013; Woodget and Austrums, 2017; Woodget et al., 2018). This methodology derives characteristic grain sizes using image texture metrics, autocorrelation or wavelet transformations. Such methods are based on regression between single parameters and grain sizes (e.g., Carbonneau et al., 2004; Warrick et al., 2009; Woodget et al., 2018), or on more novel methods based on Convolutional Neural Networks (CNN; e.g., Buscombe, 2020; Lang et al., 2021). The second approach focuses on detection and measurement of individual grain features, and is thus referred to as 'object-based'. Individual surface grains can be identified by using image thresholding and segmentation processing (e.g., Graham et al., 2005; Detert and Weitbrecht, 2012; Purinton and Bookhagen, 2019), or with most recent object detection algorithms based on CNN (e.g., Soloy et al., 2020; Mörtl et al., 2022; Chen et al., 2022). The advantages of object-based grain sizing over image statistics-based methods are (i) that the former does not require site-specific calibration and (ii) that one can derive more information from object-type data (e.g., grain arrangement). Object-type data can also be converted into grid-type data whose cell provide the local GSD.

The grain size distribution of gravel-bed rivers presents significant variations at local scale (Bluck, 1979). Many studies regarding the local GSD variability have been based on sparse manual grain size measurements combined with visual analyses (e.g., Lisle and Madej, 1992; Dietrich et al., 2006; Guerit et al., 2014). The first object-based grain sizing tools have been initially developed for applications on handheld imagery or images from pole-mounted camera (e.g., Butler et al., 2001; Graham et al., 2005). This imagery approach limits the investigation to local spatial scales. Applying object-based grain sizing methodologies to orthoimages may provide information on spatial variations in GSD that occur over distances spanning from several hundred meters to a few decimeters.

The quality of the grain size information derived from orthoimages and its potential for understanding geomorphological processes depend a great deal on the accuracy of the applied technique. Therefore, testing the latest object-based methodologies on high-resolution orthoimages is key to identify specific limitations and biases. Mair et al. (2022) evaluated the uncertainties in grain size measurements on aerial imagery with regard to the UAV-SfM approach. A performance assessment of a set of existing grain sizing routines has been conducted by Chardon et al. (2022) for applications to handheld imagery. To the best of our knowledge, no study evaluated multiple object-based techniques for applications to high-resolution orthoimages.

In this context, we evaluated three object-based grain sizing software routines (BASEGRAIN, Detert and Weitbrecht, 2012; PebbleCountsAuto, Purinton and Bookhagen, 2019; GALET, Mörtl et al., 2022) and compared their outcomes to in-field manual measurements of grain sizes and manual labelling on orthoimages. We wanted to answer the following questions. First, which automated grain-sizing software performs the best on orthoimages of a gravel-bed river? Second, which are the limitations of each tool? We then investigated how the information issued from these software routines can be used for mapping and quantifying spatial variations of surface grain sizes in a braided mountain river located in the

Swiss Alps. Global and local spatial autocorrelation statistics (e.g., Moran, 1948; Anselin, 1995) have been extensively used for assessing geographical clustering in various scientific fields—from crop and landform classifications (e.g., Maimaitijiang et al., 2020; Drăguţ and Eisank, 2012) to criminological and epidemiological investigations (e.g., Baller et al., 2001; Auchincloss et al., 2012). We wanted to explore their potential for analysing the spatial distribution of grain sizes along a river reach.

2. METHODS

The surface grain size distribution of a mountain river (Sect. 2.1) was investigated by conducting line sampling (Sect. 2.2). We designed UAV surveys to reconstruct orthoimages of the study site based on SfM algorithms (Sect. 2.3). Grain sizes were measured digitally on these orthoimages by using manual labelling (Sect. 2.4) and three object-based grain sizing methods (Sect. 2.5). These digitally measured grain sizes were compared to grain sizes measured manually in the field (Sect. 2.6). Finally, we evaluated the spatial variability of surface grain sizes over the study reach using one of the object-based grain sizing software (Sect. 2.7).

2.1 Study site

The Navisence is a mountain river located in the South-West Swiss Alps, tributary to the Rhône River (Figure 1a). This 23-km-long river drains a 257 km^2 catchment. Its main water source is the Zinal Glacier at 2300 m a.s.l. The river is hydrologically undisturbed upstream of the village of Zinal.

The study site is a 500-m-long and 60–90-m-wide river reach, located in the upstream part of a 2-km-long floodplain named "Plats de la Lée" with an average slope around 3% (Figure 1b). The Navisence flows across this alpine floodplain and develops a braided network upstream of the village of Zinal (1650 m a.s.l.). There, the catchment area is 77 km^2. A gauging station managed by a Walliser research institute (CREALP, Sion) is located downstream of the study reach, and has gathered data since 2011 (flow rates and bedload transport rates). The river has a glacio-nival hydrological flow regime, with very low flow rates in winter and high discharges in summer related to snow and glacier melting, with significant circadian variations (Travaglini et al., 2015). The typical low flow discharge is 1 m^3/s in winter, while maximum hourly discharge can exceed 25 m^3/s in summer. Over the last five years, the morphology of the braided network has been mostly impacted by a single major flood in July 2018. Regarding the sediment lithology, the surface alluvial deposits found in the river bed at the Plats de la Lée are mostly composed of metamorphic rocks (mainly orthogneiss). Therefore, sediments found on the study reach often exhibit variations of rock texture inside single grains, due to foliation or veins for example.

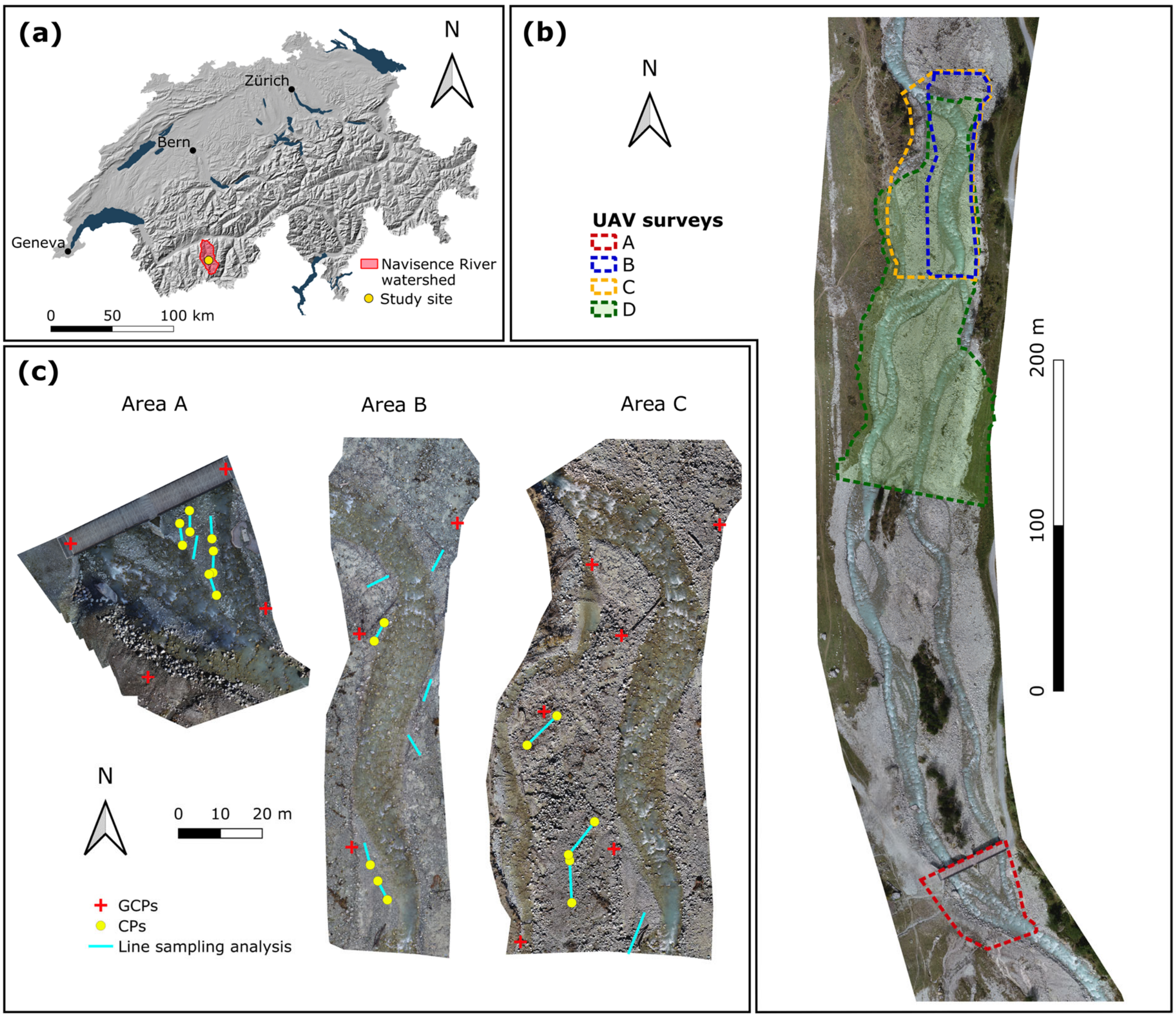


Figure 1. (a) Location of the study site and the Navisence River watershed in Switzerland (© swisstopo). (b) Location of the UAV surveys over the study reach (the river flows northwise). The river image is an orthoimage obtained from a UAV survey carried out on 13 Sep 2022 (50-m-high flight). (c) Detailed view of the three orthoimages which were reconstructed from UAV collected images, with positions of the line sampling analysis, ground control points (GCPs) and check points (CPs).

2.2 Manual measurements of grain sizes

We used the line sampling procedure proposed by Fehr (1987). This procedure has been specially devised for mountain rivers, and has thus been used in numerous field studies of hydraulics and sediment transport in mountain rivers across the Alps (e.g., Ramirez et al., 2022; Schneider et al., 2016; Konz et al., 2011; Rickenmann and McArdell, 2007; Rickenmann, 1997). There are alternatives to Fehr’s method; for instance, grid sampling protocols such as the Wolman pebble count (Wolman, 1954) and its variants are much more common than line sampling. The main advantage of Fehr's method over other techniques is that it is much easier to georeference a sample line than a grid (or a random walk) in the field and in orthophotographs. When comparing grain size estimates from field measurements and automated methods, it is thus possible to sample the same stones in the field and orthophotographs and thus, by doing so, we can benchmark methods on a fair basis. Another major advantage of Fehr’s method is that by considering all the stones crossed by the sampling line, it requires a smaller sampling space than methods based on areal or random particle sampling, which is of great interest for characterising spatial variations in grain size distribution when the riverbed involves patches of distinct grain sizes. A possible disadvantage of Fehr’s method is that its accuracy has been much less studied than Wolman-based methods (Rice and Church, 1996; Daniels and McCusker, 2010). The recommendations drawn for Wolman-based methods should, however, apply to Fehr’s method: the sampling line has to cross a large number of stones—as large as 400 according to Rice and Church (1996)—if sufficient precision in the quantile (or percentile) estimation is desired. In practice, this requirement is problematic for mountain rivers like the Navisence because of spatial variations in grain size over short distances. From a statistical viewpoint, the central limit theorem can help understand why accuracy of grain size quantiles varies as the sample size's square root as Rice and Church (1996) found in their field study, and why this result holds only when the sample is drawn from the same population, with the same mean and variance. When the grain size distribution exhibits substantial spatial variations, obtaining unbiased estimates becomes particularly difficult: (i) if the sample size is small, then it is probably representative of the local population, but estimates are inaccurate, (ii) if the sample size is large, then in principle, a higher accuracy is expected in quantile estimation, but measurements are biased. The Navisence riverbed is characterised by a typical median stone size of around 10 cm and patch length of approximately 10 m. These features led us to consider that a sample size of about 150 stones—as proposed by Fehr (1987)—provided a suitable trade-off between precision and representativeness.

We collected samples over 17 lines distributed over areas A, B and C (Figure 1c). Following Fehr (1987), we stretched a string over the dry bed-material surface to be analysed. The b-axis of all stones underneath the string was measured. Stones with a b-axis smaller than 1 cm were not considered. The stones were divided into diameter classes and the number of stones falling in each grain size interval was computed. Approximately 150 stones should be measured to ensure a good representativeness. This led us to choose sampling lengths of 5 or 10 m depending on the local grain size. The largest stones (over 10 cm) were often imbricated or clogged, which prevented a correct measurement of the b-axis. They were therefore manually extracted, which required significant effort and the use of a pickaxe. For one sample and a single operator, this procedure lasted about 1 hour. To georeference each line on geographic information systems, we measured the starting and end point positions using a pole-mounted GPS/GNSS system—Leica Zeno 20 coupled with a GG04P antenna, with real-time kinematic correction (Swipos-GIS/GEO network).

Fehr’s method involves converting the line samples into approximate volumetric-sample equivalents of the subsurface grains via empirical relations between surface and subsurface grain sizes (Fehr, 1987). The frequency-by-number grain size distribution is converted into a frequency-by-weight distribution (describing the weight fraction of each grain size interval), so that the results are comparable with standard volumetric sampling. The conversion is based on the voidless cube model (Kellerhals and Bray, 1971), which was empirically adapted by Fehr (1987). As grains whose b-axis is smaller than 1 cm are neglected during the sampling process, the cumulative frequency of the components larger than 1 cm has to be corrected to take neglected finer components into account. According to Fehr (1987) observations based on field mechanical sieving in a large set of Swiss gravel-bed rivers, 20% to 30% of the subsurface layer volume is smaller than 1 cm in diameter. Finally, the GSD is extrapolated toward the finest grain sizes. Fehr (1987) observed that for the Swiss Alps, the distribution of the fine fraction of the bed and bedload material generally follows a Fuller curve. When predicting the proportion of fine material in the GSD, Fehr assumed that the final GSD follows a Fuller curve for the undersampled finest grain sizes (see Supporting Information for the detail). We consider this tail correction to be well suited to our field site, whose bed is mostly structured by coarse particles and clogged by glacier flour (Figure 2).

Figure 2. Navisence: view from upstream of the riverbed in the southern sector of area C.

2.3 UAV surveys and structure-from-motion photogrammetry

Areas A, B and C: data acquisition

Three UAV surveys were carried out over different sectors of the study reach. The covered areas were named A, B, and C and their location is shown in Figures 1b and 1c. We conducted these surveys in order to evaluate the accuracy of digital object-based grain sizing tools on orthoimages, compared to in-field line sampling. Nine manually sampled lines were located in area A, seven in area B and four in area C. We used a DJI Phantom 4 pro and a DJI Phantom 4 pro v2 UAVs. These rotatory-wing quadcopters are equipped with a GPS for automated flights. They have an integrated camera with a 20-mega-pixel resolution. The automated flights were planned using Pix4Dcapture software (v. 4.13.1; developed by Pix4D, Lausanne, Switzerland). Images were taken vertically on a predefined trajectory, with a frontal and lateral overlap between individual images in the order of 70%. During image acquisition, the UAV stayed stationary to avoid motion blur. It then moved to the next predefined position along the grid line map. In order to obtain the best compromise between image resolution and spatial coverage, we conducted our flights at an elevation of approximately 10 m above the take-off position. As the UAVs flew horizontally, the effective flight height varied depending on ground slope and local topographic features. Ground resolution ranged from 2.9 mm/px to 3.7 mm/px. The survey C was conducted under sunny conditions, whereas the surveys over area A and B were conducted under shaded conditions, on clear days and before the sun illuminated the study reach (see Table 1). These three UAV surveys were always performed before manual line sampling. Further information about the camera parameters used can be found in Tables S1 and S2 of the Supporting Information.

Area D: data acquisition

In order to evaluate the spatial variability of surface grain sizes in the study reach, we carried out a large-scale UAV survey in a second step, covering approximately 14’000 m^2 (named area D). The flight was planned using the Map Pilot Pro software (v. 4.1.16; developed by Automotive Data Research, California). The advantage of this UAV flight planning software over Pix4Dcapture is that it makes it possible to fly at a constant height above a digital elevation model (DEM) generated from NASA's Shuttle Radar Topography Mission (resolution of 30 m/px). Therefore, images could be collected at similar flight height (about 10 m) as the approximate floodplain topography was considered. The survey was carried out using a DJI Phantom 4 pro. To avoid motion blur in the images, we selected a fast shutter speed (see Table S2 of the Supporting Information for specific camera parameters). The survey was carried out on a clear sky day when the river reach was shaded. Images were taken along a single grid line map at a rate of one image per second and the flight speed of the drone was set at 2 m/s. The images had an overlap in the order of 60% in flight direction and a lateral overlap of approximately 70%. The other characteristics of this survey are summarized in Table 1.

Table 1. Summary of the UAV surveys (GCP = Ground Control Points and CP = Check Points).

Area name	Area size [m²]	Date	Weather	UAV model	Number of images	Number of tie points	Number of GCP|CP	Ground sampling distance [mm/px]
A	2’500	5 Oct 2022	Shady, clear sky	DJI Phantom 4 pro	232	6’070’479	4 | 9	2.9
B	3’450	12 Oct 2022	Shady, clear sky	DJI Phantom 4 pro	223	6’053’898	3 | 5	3.7
C	5’850	26 Oct 2022	Sunny	DJI Phantom 4 pro v2	419	12’511’931	6 | 6	3.3
D	14’350	11 Nov 2022	Shady, clear sky	DJI Phantom 4 pro	787	17’680’434	8 | 21	3.1

Auxiliary georeferenced points

Ground control points (GCP) were distributed over the surveyed areas to constrain more accurately the SfM photogrammetric reconstruction and to assess errors. The GCP were marked with paint, and their position was measured using the same GPS/GNSS system described in Section 2.2. This provided a horizontal positioning accuracy close to 1 cm and a vertical accuracy in the 2–4-cm range. The GCP coordinates were measured in the CH1903+/LV95 coordinate system (EPSG 2056). During the SfM reconstruction preprocessing steps, we located the GCP position on the images. The start and end points of the manually-sampled lines served as independent check points to evaluate the accuracy of the reconstructed orthoimages. Some start and end points could not be located with certainty on the basis of the photographs taken in the field (14 out of 34), and were therefore not considered check points. Eight GCPs were defined in area D. Additional check points were defined on the basis of three orthoimages produced from 50-m-high flights covering the entire “Plats de la Lée” (UAV surveys pertaining to a different study). Twenty-one fixed features were identified on these three orthoimages and in the orthoimage of area D (see Table 1). The mean coordinates of these features labelled in the three orthoimages served as check points. The resulting orthoimage positioning errors are given in Table 2.

Table 2. Quality assessment of Structure-from-Motion photogrammetry results. Mean error (ME) and standard deviation of error (STDEV) on ground control points and check points after bundle block adjustment.

Area name	Ground Control Points			Check Points		
	X [cm] ME | STDEV	Y [cm] ME | STDEV	Z [cm] ME | STDEV	X [cm] ME | STDEV	Y [cm] ME | STDEV	Z [cm] ME | STDEV
A	0.0 | 0.8	0.0 | 0.9	0.0 | 1.0	0.2 | 1.5	-2.8 | 2.3	0.3 | 2.8
B	0.0 | 0.8	0.0 | 0.5	0.0 | 0.0	-0.9 | 1.5	-1.8 | 2.1	-4.2 | 2.0
C	0.0 | 1.0	0.0 | 0.9	0.0 | 1.2	0.8 | 1.8	0.7 | 1.7	2.8 | 1.4
D	0.0 | 7.3	0.0 | 5.4	0.0 | 1.4	-0.1 | 4.1	2.1 | 7.3	2.5 | 3.8

Data processing

Geo-referenced orthoimages were obtained by processing the images with the Pix4Dmapper software (v. 4.8.0; Pix4D, Lausanne, Switzerland). By combining SfM photogrammetry and multi-stereo view algorithms, the software reconstructs the three-dimensional (3D) surface topography. In the SfM framework, the 3D positions of a large set of features automatically extracted from images are retrieved, simultaneously with camera positions and orientations by iteratively solving a highly redundant system of triangulation equations (Westoby et al., 2012). This method provides a point cloud, which can then be converted into a DSM and an orthoimage. The main parameters used in Pix4Dmapper can be found in Table S3 of the Supporting Information.

2.4 Digital manual labelling

Manual labelling was performed by a single operator on the orthoimages using the QGIS software (v. 3.22). We manually drew a polygon on all visible grain features intersected by the georeferenced lines. This labelling operation took approximately 10–15 minutes per line. We measured the b-axis of all labelled grains by automatically fitting an ellipse to each feature (fitting based on the second central moment of the object geometry). The detailed ellipse fitting procedure is described in the Supporting Information. The b-axis corresponds to the minor axis of the fitted ellipse. This procedure for extracting the b-axis of each labelled grain is similar to the ellipse fitting procedure used by the object-based grain-sizing methods described in Section 2.5. Finally, we derived the GSD from the b-axis of the identified grain features for each line using Fehr's (1987) method implemented in a Python script.

2.5 Description of selected object-based grain sizing tools

In this section, we describe the internal frameworks of the three object-based grain sizing tools under investigation. We specify how each tool was implemented to derive grain size distributions that are comparable with those based on in-field manual samples. Figure 3 shows an example of detected grain features along a line using the different methodologies.

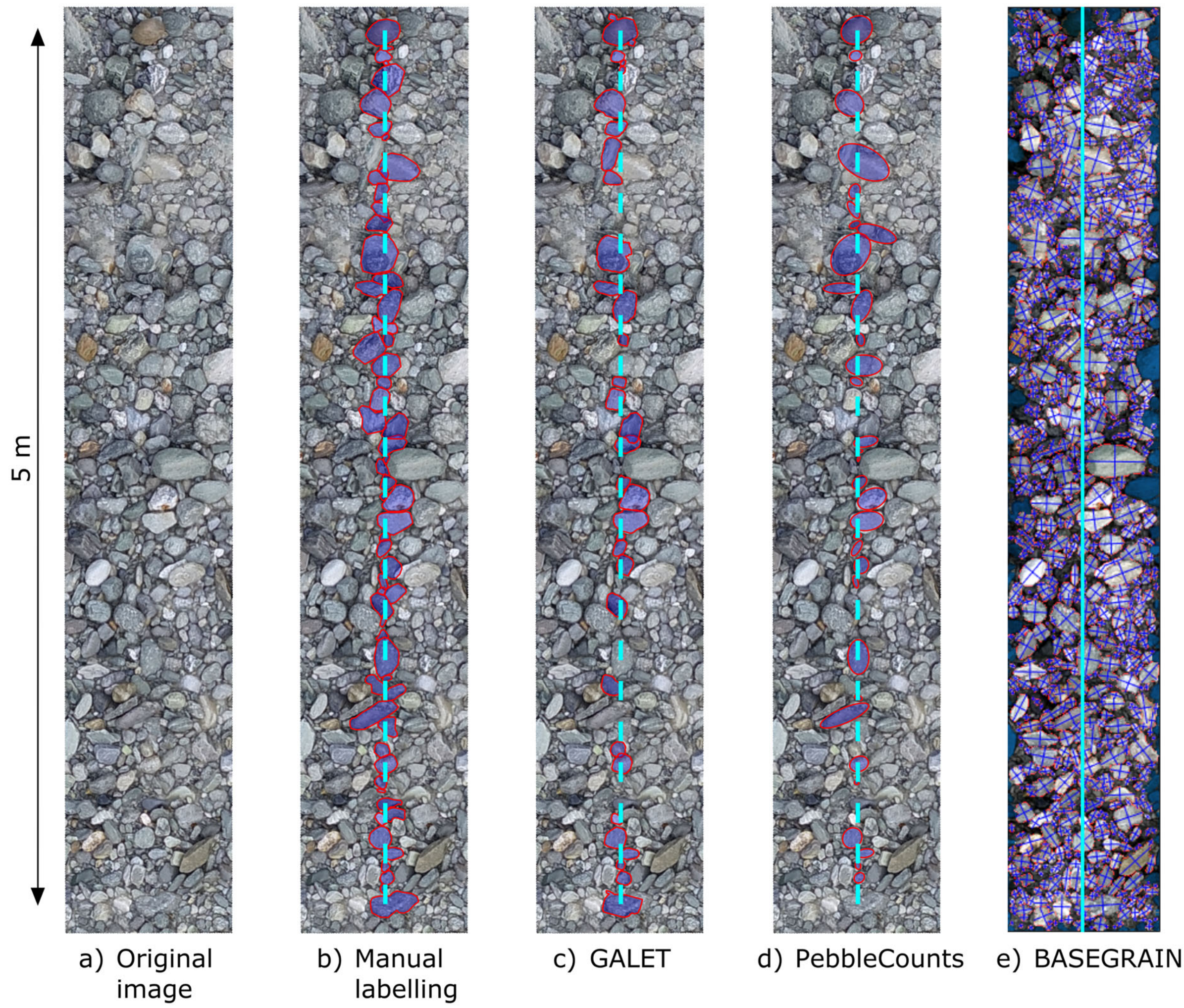


Figure 2. Example of digital line sampling with manual labelling and the software routines investigated. The blue line corresponds to the location of the georeferenced line. The grain features displayed for manual labelling and GALET are the original grain feature polygons and not the fitted ellipses. The BASEGRAIN image originates from the software GUI.

BASEGRAIN

We used BASEGRAIN (v.2.3), which is a free access MATLAB-based method developed by Detert and Weitbrecht (2012). It performs individual grain segmentation on digital top-view photographs in five preprocessing steps. Three out of five steps require supervised parameter tuning to optimise performance. Ellipses are then fitted to the detected objects, and the minor axis is considered as the b-axis of the grain feature.

Orthoimages were cut out into image tiles corresponding to the bed surface patches where the line sampling was conducted. This splitting was required because BASEGRAIN is not able to handle georeferenced orthoimages. The image tiles were rotated in order to position the line vertically. The BASEGRAIN processing was much influenced by the variations in colour and texture of the Navisence sediments, thus no unique set of parameters allowed for optimised object detection for all image tiles. Therefore, object detection had to be performed individually for each image tile when using BASEGRAIN. We tuned the different parameters until a visually optimal segmentation of the grains was obtained. No postprocessing was applied to the detected objects. The virtual sampling line implemented in BASEGRAIN was placed so as to match the position of each field sampling line. We extracted the dimensions of all the grains detected and intersected by the line in BASEGRAIN. We then tallied grains using the same grain size intervals as those used for manual sampling, and computed the GSD according to Fehr’s (1987) method.

GALET

GALET is a deep-learning image segmentation model for grain size analysis developed by Styx4D (Bourget-du-Lac, France) and presented in Mörtl et al. (2022). The CNN model implemented in GALET was trained using a dataset generated by a technique based on automated image creation. Mörtl et al. (2022) used manually-cropped grain images and synthesised artificial grain images to generate labelled training images. During the grain detection and segmentation steps in GALET, orthoimages are split into 512 or 1024 pixel-large tiles. The software performs grain detection and estimates the shapes of overlapping grains. A final shapefile is produced in which all detected instances are vectorised.

The entire GALET segmentation process was applied on each orthoimage. The routine detected the grain features and provided the corresponding shapefiles. We measured the b-axis of all detected grains by automatically fitting an ellipse using the same method as the one utilised for manual labelling. Digital line sampling was performed on these shapefiles in QGIS, as the positions of the field sampling lines were georeferenced. We classified the vectorised grains that intersected the georeferenced lines according to their b-axis. The same grain size intervals used for manual sampling were considered and the GSDs were computed by using the Fehr (1987) method.

PebbleCountsAuto

PebbleCounts is an open-source Python-based algorithm, developed by Purinton and Bookhagen (2019). Here we used, its highly-automated version named PebbleCountsAuto. It is an image segmentation method that performs individual grain detection. Resulting grain features are measured via ellipse fitting.

As PebbleCountsAuto requires significant computing time to process entire orthoimages (Purinton and Bookhagen, 2021), grain feature detection was performed on image tiles cut out from the orthoimages (similarly to the procedure used with BASEGRAIN). These tiles corresponded to the locations where in-field line sampling was performed. PebbleCountAuto only required us to manually tune the threshold level of Otsu's threshold matrix. This parameter was tuned for each processed image tile in order to obtain a visually optimal segmentation of grain features. The default parameter defining the minimum area in pixel for a feature to be considered a grain was modified and set at 23 pixels in order to be consistent with the same parameter defined in BASEGRAIN—value based on the limit of grain feature detectability in images (see Graham et al., 2005). The model uses a size cut-off criterion to discard grains whose b-axis is too small. By default, the cut-off value is set at 20 px, but we decreased it to 3 px so that all grains with a b-axis exceeding 1 cm could be considered. PebbleCountsAuto allows one to work with georeferenced orthoimages. Therefore, information about detected grain features such as northing and easting coordinates in an UTM coordinate system, major and minor axis of the fitted ellipses and their orientation could be exported as text files. This data was used to reconstruct the detected features as georeferenced ellipses in QGIS. Digital line sampling was then performed by computing the GSD from grain features intersected by each georeferenced line.

2.6 Accuracy evaluation

The grain size percentiles obtained by manually labelling images and those obtained by applying the three grain-sizing tools were compared with the grain sizes retrieved by on-field manual line sampling. The comparisons were done by normalising all digitally obtained grain size percentiles by their corresponding in-field manually measured grain size percentiles as follows:

$$\mathrm{d}_{\mathrm{norm}} = \frac{\mathrm{d}_{\mathrm{digital}}}{\mathrm{d}_{\mathrm{manual}}}$$

where $d_{digital}$ corresponds to the digitally obtained grain size percentiles, while d_{manual} corresponds to the grain size percentiles obtained from in-field manual sampling. Therefore, if the digital measurements were accurate, d_{norm} should be close to unity.

We considered the Normalised Root Mean Square Error (NRMSE) to quantify the errors of digitally-based grain size percentiles in terms of the corresponding fraction of the mean grain size percentiles obtained from in-field manual sampling. This error metric allowed us to directly compare the accuracy of the model estimates for different grain size percentiles, even if the grain scales were different (like for d_{16} and d_{84} values which may not be of the same order of magnitude). For each grain size percentile, the NRMSE was calculated as follows:

$$\mathrm{NRMSE} = \frac{\sqrt{\sum_{i=1}^{n}(y_i - x_i)^2}}{\sqrt{n} \cdot x_{mean}}$$

where y_i is the value of the digitally estimated grain size percentile on sample i, x_i is the value of the manually measured grain size percentile on sample i, x_{mean} is the mean value of the manually measured grain size percentiles and n is the number of line samples—n = 17 in this study.

2.7 Spatial variability of grain sizes

The spatial variability of surface grain sizes in the study reach was investigated by considering the spatial distribution of the dataset of grain geometries detected by the GALET software on the orthoimage covering the area D reach. The objective was to determine whether grain sizes were randomly distributed over the river bed, or whether some spatial arrangements could be statistically identified in the grain-size samples. We investigated this spatial variability (i) in a regular grid of d_{50} estimates computed from the detected grain features, and (ii) directly in a set of discrete grain features.

The d_{50} estimates were computed on 2 × 2 m grid cells. This cell size was chosen as a trade-off between local GSD representativeness (that is, the cell was large enough to contain approximately 100 stones in most cells) and spatial variability (that is, the cell was small enough to exhibit grain-size variations at the bar scale). Grid cells including water were not considered, as GALET is generally unable to detect underwater grain features. We computed the GSD from the grain features detected by GALET, by considering all detected stones whose centroid was located in the cells. This sampling procedure corresponds to areal sampling (Kellerhals and Bray, 1971). According to the voidless cube model, the factor α for converting the frequency-by-number into equivalent volumetric frequency-by-weight is equal to 2 for areal sampling (Church, McLean and Wolcott, 1987). We applied a correction factor to the finest fraction, as described in Fehr’s method (which assumes that 25% particles are smaller than 1 cm, and that the lower end of the GSD can be captured using the Fuller curve). We extracted the d_{50} value from the GSD on each cell.

Metrics of spatial autocorrelation were computed to investigate whether the grains are distributed randomly in space according to their size, and thus to evaluate the overall clustering tendency. Spatial clusters are defined by Knox (1989) as "geographically bounded groups of occurrences of sufficient size and concentration to be unlikely to have occurred by chance." A commonly used indicator of clustering is Moran’s spatial autocorrelation index (referred to as Moran’s I; Moran, 1950; see Getis, 2010). It summarises the correlation between the value of one spatial unit and the mean value of its neighbouring units, via a spatial weight matrix. For the maps of d_{50} estimates, the neighbours (and thus spatial weights) of each cell were defined according to a queen contiguity, meaning that all cells sharing a common border or one vertex are considered neighbours (i.e. the 8 cells surrounding each cell). For the grain-feature shapefile, the spatial weights were set for the 100 nearest neighbours of each feature. This number corresponds to the size of typical surface samples (Bunte and Abt, 2001), and thus allows to assess the spatial autocorrelation at a grain sample scale.

Moran's I varies between -1 and 1, and takes values close to 0 when there is no spatial autocorrelation (i.e. individual variables are independent from the mean variable of neighbouring units). It tends towards 1 or -1 when there is strong positive or negative spatial autocorrelation, respectively. The variables considered here were the d_{50} estimates (grid-type data) and the b-axes of the sample of discrete grain features. By plotting the standardised variables (i.e., variables rescaled to have a mean of 0 and a standard deviation of 1) against the mean of standardised neighbouring variables in a so-called Moran scatter plot, Moran's I can be computed as the least square slope of the regression through the origin (Anselin, 1996). This procedure for computing Moran's I is equivalent to the formal definition of I, but also enables the visualization of spatial associations in the dataset (Anselin, 1996). The statistical significance of Moran's I was evaluated by testing the null hypothesis that Moran's I is equal to 0, meaning that d_{50} estimates or grain-sizes would be distributed randomly across the space. The reference distribution of Moran's I under the null hypothesis was derived by applying a Monte-Carlo method. The attributes of all individuals in the dataset were randomly moved over the locations and Moran's I was computed. This process of randomisation followed by Moran's I computation was performed 999 times, to infer the probability density function of Moran's I under the null hypothesis. Finally, we assessed the statistical significance through the pseudo p-value (Anselin, 1995). We compared Moran's I of the original dataset to Moran's I density function obtained from the Monte-Carlo simulations and rejected the null hypothesis if the p-value was lower than a significance level of 0.05.

Moran's I provides information about the overall trend towards clustering in the spatial distribution of grain sizes; it does not give any indication about cluster location. To evaluate the strength variations in spatial autocorrelation with location, we conducted an analysis of the Local Indicator of Spatial Association (LISA; Anselin, 1995). Following Anselin (1995), we decomposed Moran's I into local coefficients for each individual observation. The local Moran value I_i for each spatial unit is computed as follows:

$$I_i = \frac{(x_i - \bar{x}) \cdot \sum_{j=1}^{n} w_{ij}(x_j - \bar{x})}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

where x_i is the value of unit i (i.e. its d_{50} value or its b-axis size), x_j is the value of unit j, $\bar{x}$ is the mean value of all units, w_{ij} is the weight that defines the relationship between units i and j ($w_{ij} = 1$ if j is a neighbour of i, $w_{ij} = 0$ otherwise) and n is the total number of spatial units. We assessed the statistical significance of each spatial unit's spatial correlation with its neighbours (i.e. local Moran value), similarly to what was done for the global Moran I. We computed a density function of the local Moran for each spatial unit by running a Monte-Carlo simulation—involving 999 random permutations (significance level of 0.05). In the resulting LISA map, data were partitioned into five topological relationships: 'high–high', 'low-low', 'low–high', 'high-low', and 'not significant'. 'High–high' and 'low-low' indicated a positive spatial autocorrelation, which implies the presence of clustered high (respectively low) values. 'Low–high' and 'high-low' indicated a negative spatial autocorrelation, implying a low value with a high-value neighbourhood (respectively a high value with a low-value neighbourhood). 'Not significant' indicated the absence of significant spatial autocorrelation. All spatial statistics were computed using the GeoDa software (Anselin et al., 2010).

3. RESULTS

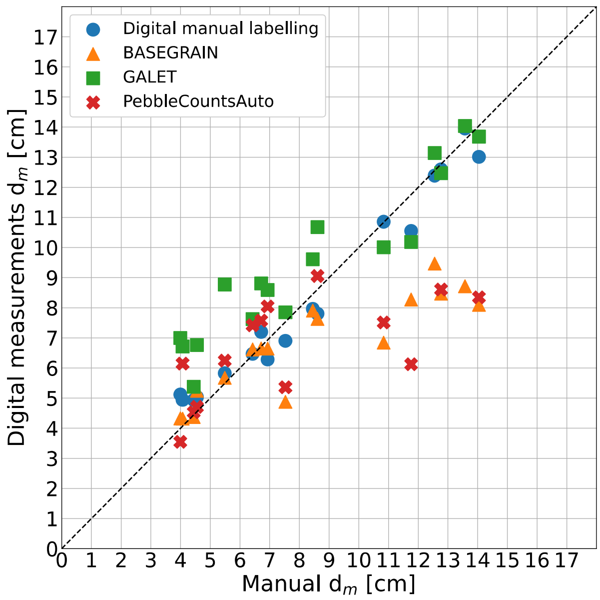
Section 3.1 covers the results regarding the accuracy of digital grain sizing methods. Section 3.2 presents the outcomes of the analysis of the grain-size spatial variability.

3.1 Digital grain sizing accuracy

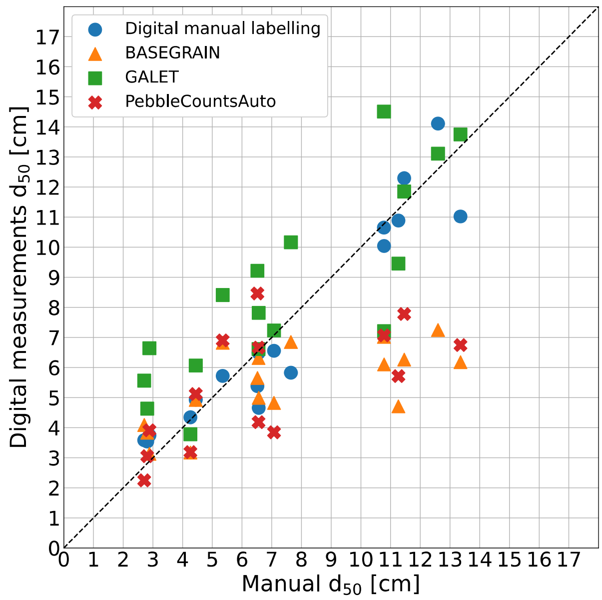
Pairs of digitally and manually measured characteristic grain sizes are presented in Figure 4 for the set of Fehr line samples collected during the field campaign. Grain size estimates from manual in-field line samples were regarded as the reference “ground-truth” values when comparing the performance of the different methods. Overall, the data pairs corresponding to digital manual labelling and their associated in-field manual measurements proved to be mutually consistent. Grain size estimates from digital manual labelling closely follow the 1:1 trend when plotted against their in-field ground-truth counterparts (Figure 4). The grain size estimates issued from GALET follow the 1:1 trendline with a particularly-good agreement on the highest half of the grain size domain for the d_m, d_{50} and d_{84} cases. However, the GALET estimates were often larger than the manual ones on the lowest half of the grain size domain (Figure 4a, 4b and 4d).

PebbleCounts grain size estimates included three outliers. For the sake of readability, the outliers were not plotted in Figure 4, as they differed strongly from manually measured grain sizes for the d_m and d_{84} grain sizes (their values ranged from 20 cm to 28 cm for the d_m grain size, and from 42 cm to 62 cm for the d_{84} grain size, see Figure S2 in Supporting Information). These outliers arose because a large grain feature (b-axis > 40 cm) was detected in each of the three concerned samples. Such large features resulted from the undersegmentation of sediment patches in PebbleCounts.

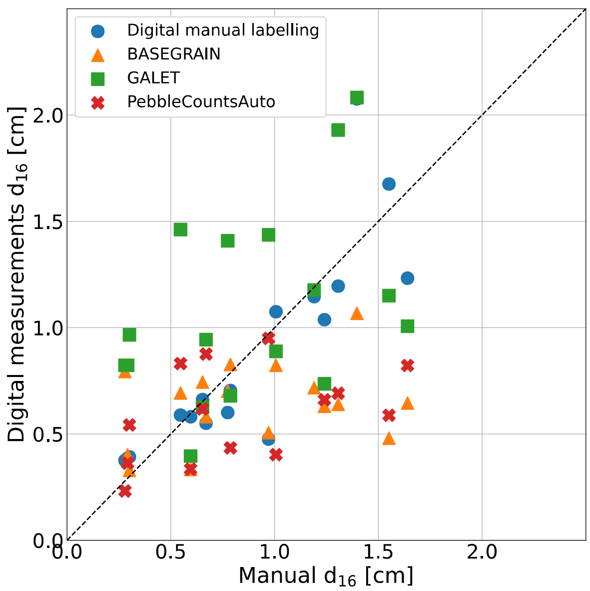
When ignoring the above outliers, we found that PebbleCounts and BASEGRAIN grain size estimates were often located under the identity line in Figure 4. Particularly, the largest manually measured values were systematically underestimated by BASEGRAIN and PebbleCounts routines for all characteristic grain sizes.



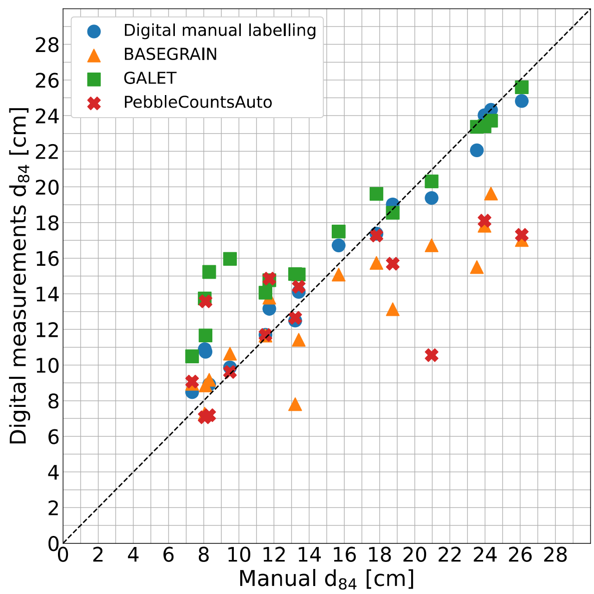
a) d_m measurements comparisons



b) d_{50} measurements comparisons



c) d_{16} measurements comparisons



d) d_{84} measurements comparisons

Figure 4. Characteristic grain size values from digital measurement procedures plotted against on-field manually measured characteristic grain sizes, for each of the 17 line sampling analysis performed. The black dashed line corresponds to the 1:1 trend.

The normalised grain size percentiles are presented in Figure 5. The median normalised grain size percentiles derived from manual labelling are close to unity for all grain size percentiles. This indicates a good match with grain sizes issued from in-field manual sampling. The normalised grain size percentiles computed by GALET were frequently in excess of the grain size values derived from in-field manual measurements (median value above unity). This overestimation was particularly pronounced for grain size percentiles smaller than d_{40}. Normalised grain size percentiles obtained from BASEGRAIN and PebbleCounts (whose outliers were not considered) underestimated grain size percentiles on average. This underestimation was less severe for larger grain size percentiles. If we look at the q_{25} and q_{75} quartiles in Figure 5, the object-detection software routines provide normalised grain size percentiles that are more scattered in the lower half of the grain size percentiles (i.e. between d_5 and d_{50}) than in the upper half, with a minimum scatter reached around the d_{80} grain size. This trend was also observed for manual labelling, but it was less pronounced—the interquartile range of the normalised grain size percentiles was relatively low compared to the object detection software estimates.

The evolution of the NRMSE as a function of the grain size percentile is presented in Figure 6. This error metric indicates that the digital measurement procedures were most accurate around the d_{84} grain size percentile, except for PebbleCounts when its outliers were considered. Grain sizes computed from digital manual labelling showed the lowest

NRMSE. Concerning the software routines, they exhibited mutually-similar error values for grain size percentiles between d_5 and d_{40}. Between the d_{40} and d_{90} grain sizes percentiles, GALET showed the lowest errors, whereas BASEGRAIN and PebbleCounts (without outliers) exhibited comparable higher NRMSE values. The NRMSE of grain size estimates of PebbleCounts was significantly reduced when removing the outliers from the error metric computation.

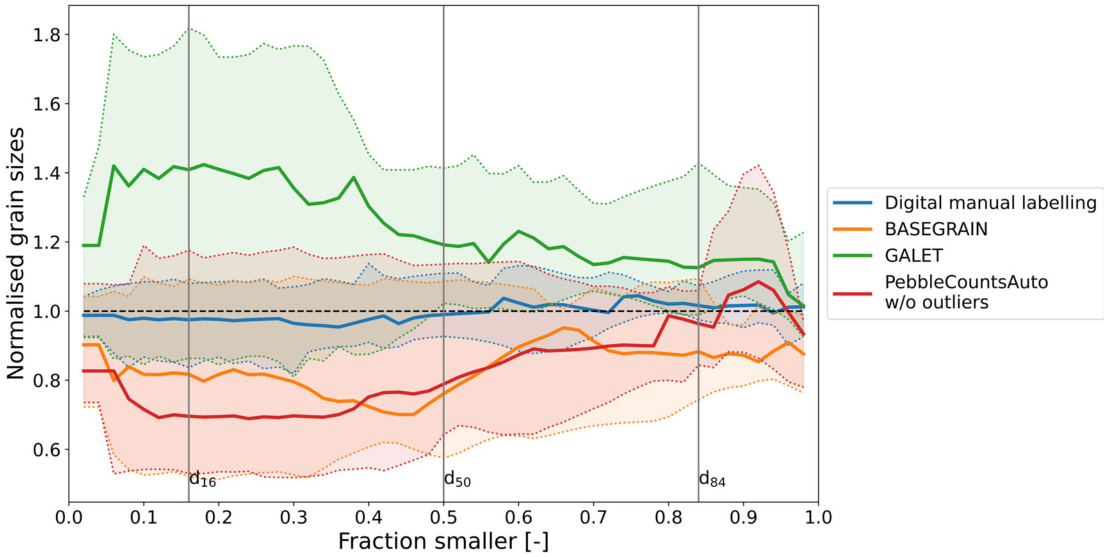


Figure 5. Median normalised grain size percentile of each digital measuring procedure (thick line). The grain size percentiles were normalised by the in-field manually derived grain size percentiles ($d_{digital}$ / d_{manual}). The q_{25} and q_{75} quartiles of the normalised digital estimates are represented by the dotted lines, meaning that 50% of the normalised estimates are located within the coloured area. Key grain size percentiles (i.e., d_{16}, d_{50} and d_{84}) are indicated by vertical lines.

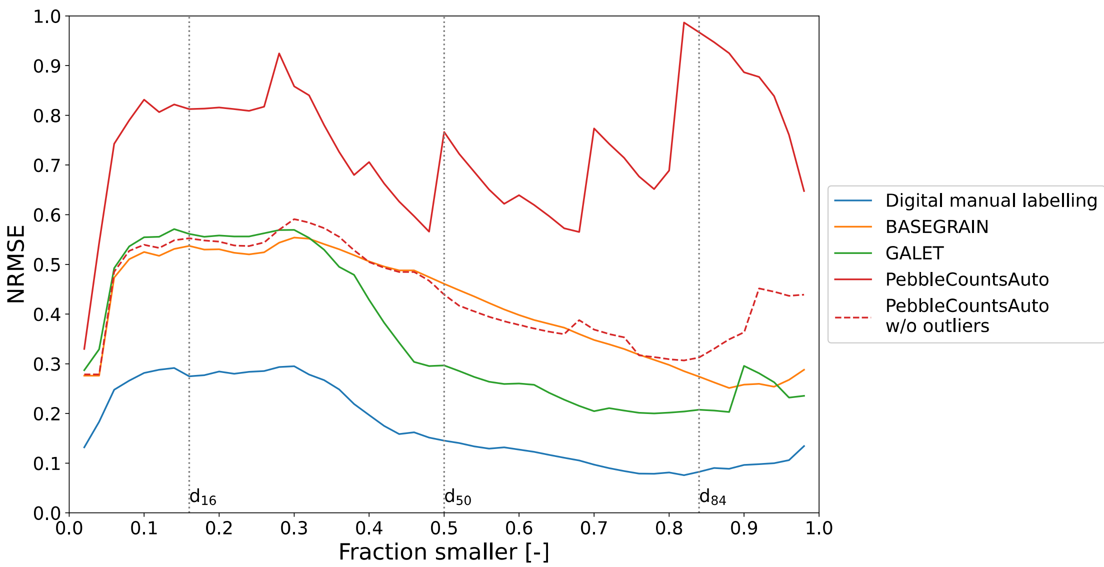


Figure 6. Normalised Root Mean Squared Error (NRMSE) of each digital measuring procedures compared to in-field manual sampling, for each grain size percentile. Key grain size percentiles (i.e., d_{16}, d_{50} and d_{84}) are indicated by vertical dotted lines.

3.2 Spatial variability of grain sizes

Given the performances of the GALET software and its capability to process large orthoimages, we chose this tool to analyse the spatial variability in surface grain sizes within the area D reach. Figure 7 shows how heterogeneous the spatial distribution of surface grain sizes was over area D. In particular, the d_{50} estimates revealed important spatial variations in the surface grain sizes over the study reach, with local d_{50} estimates spanning from 2 cm to 40 cm (Figure 7b). Visual assessment indicated that the grain size pattern identified in the map of d_{50} estimates is consistent with the spatial distribution of grain sizes observed in the orthoimage.

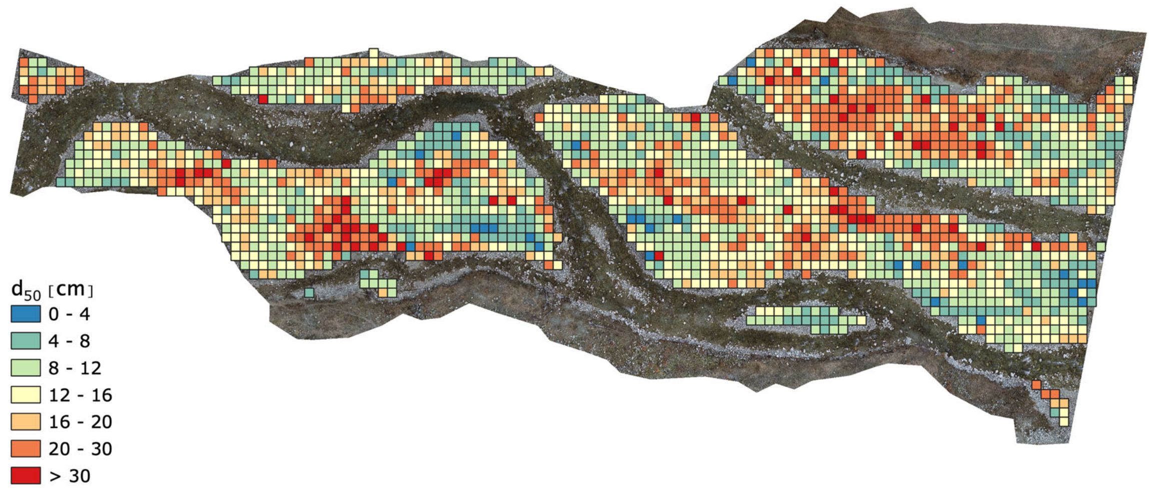
The Moran scatter plot (Figure 8) indicates that the spatial distribution of d_{50} estimates in 2 × 2 m cells was characterised by a positive spatial autocorrelation. The test of significance suggested that the Moran's index value I = 0.41 is statistically significant. The null hypothesis (stating that d_{50} estimates are randomly distributed across the space) was strongly rejected, as the computed pseudo-p-value resulted equal to 0.001 (meaning that no Moran's I computed via random permutations is as large as the observed Moran's I). Therefore, the analysis of the Moran's I confirmed that sediments are sorted into patches of similar grain sizes on the study reach.

The local indicator of spatial association (LISA) allowed us to locate statistically-significant clusters of large (high-high) and small (low-low) d50 estimates (Figure 7c). Clusters are sets of contiguous units that exhibit significant and similar spatial associations (Anselin, 1995). Relatively few cells showed negative spatial autocorrelation (i.e. significant low-high or high-low spatial associations with their neighbours). Among the clusters identified on the LISA map, three are highlighted in Figure 7c, which makes it possible to compare their location with that of the clusters identified at the grain scale in Figure 10. Note that given their size and morphology, these clusters are generally referred to as “patches” in the gravel-bed literature (e.g., Lisle and Madej, 1992; Dietrich et al., 2006). We use the term “cluster” thereafter for consistency with the definition used in the field of spatial statistics given by Anselin (1995).

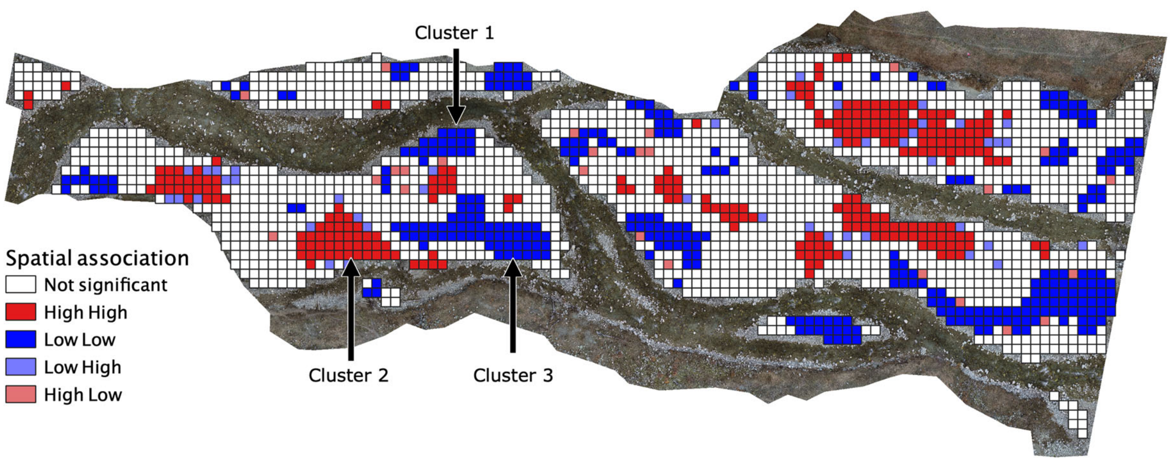
To study the spatial variability in the grain sizes within a single geomorphological unit at the grain scale, we selected the detected features located over a gravel bar in area D (Figure 10). Moran’s I computed from the resulting dataset indicates that the spatial grain size distribution was characterised by a positive autocorrelation (Figure 9). The Moran’s index value I = 0.12 was statistically significant, as the null hypothesis of random spatial distribution of grain sizes was strongly rejected (pseudo-p-value = 0.001). The positive Moran’s I value indicated that the surface stones were generally clustered according to their dimensions in the surveyed river reach. The large number of grain features presented in the Moran scatter plot and the relatively weak correlation makes it difficult to visually detect a clear trend in the data scatter around the linear regression line (Figure 9). The smallest standardised b-axis values are located within one standard deviation from the origin. This indicates that the grain-size distribution was largely skewed towards fine particles (approximately 70% of the detected particles were smaller than the mean b-axis value of 6.2 cm). We recall that the amount of information related to the finer grain sizes was hindered by the image resolution. The smallest particles which could be detected by GALET were of 2 cm in diameter; this limitation imposed a lower bound in the data scatter.

The cluster map generated from the LISA allowed us to visualise different grain size clusters (Figure 10). A complex pattern of different surface grain sizes was observed on this bar, where the clusters 1, 2 and 3 identified in Figure 7c are located. It is worth recalling that LISA outcomes and cluster identification could be influenced, to a certain degree, by the characteristics of the data sample considered (e.g., size and internal variability). The clusters identified from the LISA map computed from d_{50} estimates on a grid matched reasonably well—in location and extent—those identified from the LISA analysis based on discrete grain features (Figure 10). This good agreement means that, in our case, the dataset of samples located on the single geomorphological unit considered were highly representative of the grain size distribution over the entire area D.

a) Orthoimage of the area D.



b) GALET d_{50} estimates.



c) Local indicator of spatial association of d_{50} estimates.

Figure 7. Maps of the spatial distribution of d_{50} estimates. (a) Orthoimage of area D on 11 Nov 2022, with discharge of 1.2 m^3/s. The water flows northwise. (b) d_{50} estimates based on detected grain features by GALET in 2 × 2 m grid cells. (c) Cluster (i.e. patches) identification using the local indicator of spatial association (based on the local Moran's I), three clusters are indicated for further discussion.

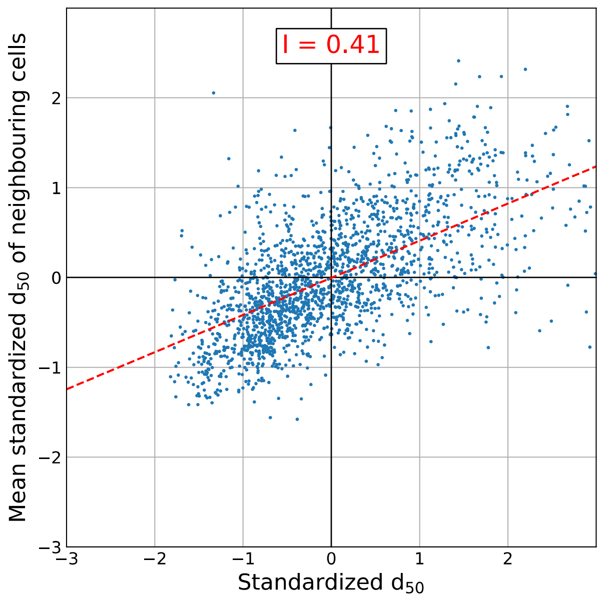


Figure 8. Moran scatter plot for the measure of the overall size clustering tendency of the grid d_{50} estimates. The mean standardised d_{50} estimates of neighbouring cells are plotted against the standardised d_{50} estimates. The global Moran's I corresponds to the slope value of the regression line (dashed red line).

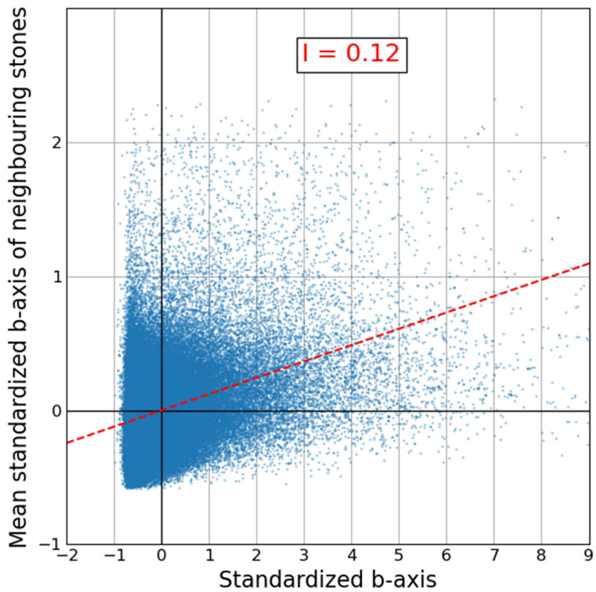
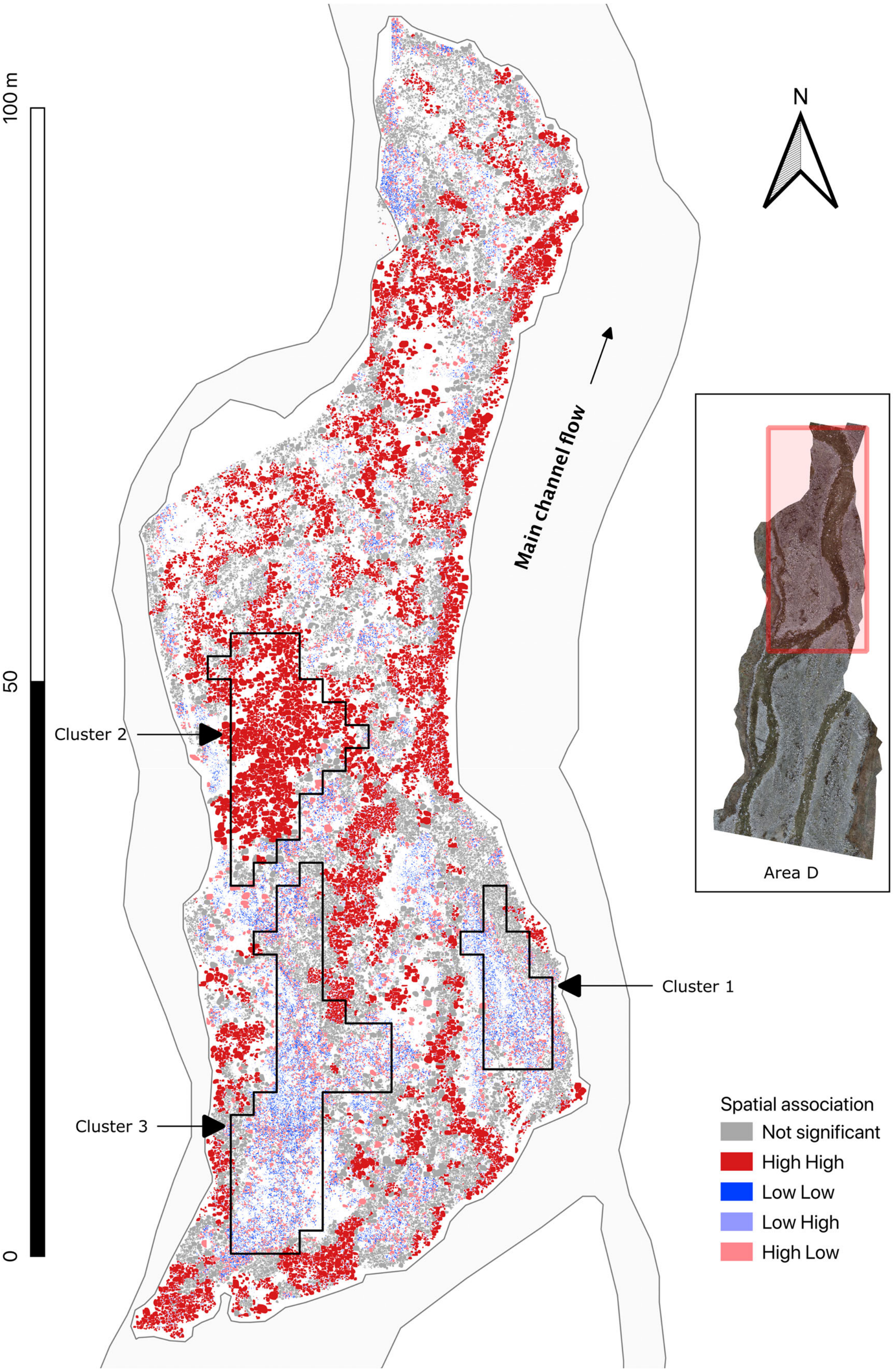


Figure 9. Moran scatter plot for the measure of the size clustering tendency of grain features. The mean standardised b-axis estimates of the 100-nearest neighbouring stones are plotted against the standardised b-axis of the 125'000 detected stones by GALET. The global Moran's I corresponds to the slope of the regression line (dashed red line).

744



745 **Figure 10.** Particle-size clusters (i.e. patches) identification using a local indicator of
746 spatial association (local Moran's I). The region of interest is a single gravel bar located
747 in area D. The contours of the clusters 1,2 and 3, as identified from the LISA map for d_{50}
748 data, are plotted for comparison (solid black lines).

4. DISCUSSION

The accuracy of digital manual labelling and of object-based grain sizing methods compared to in-field line sampling is discussed in Sections 4.1 and 4.2, respectively. Section 4.3 addresses how spatial statistics of detected grain features can be used to describe the spatial organisation of surface grain sizes.

4.1 Digital manual labelling accuracy

The characteristic grain size values derived from manual labelling of orthoimages showed a great similarity with the values derived from in-field manual sampling (Figure 4 and 5). This similarity contrasts with previous studies on manual labelling of individual grains in photographs, which found that grain sizes were generally underestimated compared to in-field sampling results (e.g., Adams, 1979; Ibbeken and Schleyer, 1986; Church, 1987; Garefalakis et al., 2023). These authors linked this underestimation to the partial information accessible in photography-based grain sizing methods, since only the exposed part of grains is visible in nadir photographs. Partial burying of grains, grain imbrication or foreshortening of grains due to the angle of the photograph can lead to underestimate the true grain sizes (Graham et al., 2010).

In our study, grain size percentiles derived from digital manual labelling did not suffer from underestimation, most likely because it was difficult to identify smallest particles (b-axis < 2 cm approximately) owing to the orthoimage resolution or to their location in-between coarser particles.

The weak detection of finest particles resulted in different calibrations of the Fuller curves that describe the lower end of the GSDs and probably led to higher NRMSE values for grain size percentiles in the d_{10}–d_{40} range (Figure 6). In our samples, percentiles smaller than d_{10} corresponded to grain sizes smaller than 1 cm and no grain-sizing method (including in-field sampling) provided direct measurements for such small grains. Therefore, under d_{10}, the Fuller interpolation in the Fehr method completely determined the grain-size percentiles, made the GSD tails mutually-similar and provided lower NRMSE values. The undersampling of fine grains likely counterbalanced the size underestimation for the largest particles, thus providing average grain size estimates that were similar to those derived from in-field manual sampling. It is worth mentioning that manual labelling was performed by a single operator and we did not investigate how the results may differ depending on the operator.

4.2 Accuracy of object-based grain sizing methods

Comparing the three software routines (BASEGRAIN, GALET and PebbleCountsAuto) revealed differences in accuracy and limitations. GALET tends to overestimate in-field grain sizes issued from line sampling. The main explanation for this phenomenon is that GALET did not detect the smallest grains (b-axis < 2–3 cm). Mörtl et al. (2022) noted that the resolution of orthoimages determines the smallest detectable grain size by GALET. The resolution of the generated orthoimages (approximately 0.3 cm/px) was likely too low for the detection of the smallest grains. Fine grained samples (with d_m smaller than 7 cm, Figure 4a) were therefore particularly affected by the absence of small grains in GALET grain size estimates and thus led to overestimated values. To reduce the detection limit for small grains, shorter ground sampling distances would be required.

Visual inspection of the grain features detected by GALET suggested that the software performance was not affected by different rock texture patterns inside individual grains. The CNN training dataset used by Mörtl et al. (2022) is probably well suited to applying the routine to the Navisence bed images. A significant number of grains was not detected in the GALET routine. Overall, GALET detected 40% less grain features, regardless of their size, along the lines compared to the manual labelling conducted by a human operator. The largest grains (b-axis > 20 cm) sometimes appeared over-segmented by a vertical or horizontal line. This feature splitting was caused by the edges of the finite-size moving window used for grain feature detection in GALET. This issue did not arise with large grain features along the lines, but it could lead to a size underestimation for some of the largest stones present in the river bed.

BASEGRAIN and PebbleCounts produced similar results, as both software routines generally underestimated characteristic grain sizes. For three sampling lines, PebbleCounts outcomes were affected by several feature merging occurrences and missed grain detections (the outliers mentioned in Section 3.1). These errors were likely caused by glacial flour, which partially covered pebbles. This led to a large overestimation of grain size percentiles for these lines. The grain sizes of the other lines were generally underestimated because the largest stones were often not detected, which may be due to the presence of intergranular textures that prevented optimal edge detection. In addition, direct sun illumination on orthoimage C caused size underestimation of the detected stones, as the shaded grain faces were not included in the detected object boundaries. Finally, we observed that the grain masks identified by PebbleCounts were generally smaller than the apparent size of stones in the orthoimages. This shortcoming in PebbleCounts led to grain size underestimation. PebbleCounts showed a poor detection rate along the lines on the orthoimages, as the number of grain features detected is 62% smaller than that resulting from manual labelling.

Concerning BASEGRAIN, rock-texture variations inside single grains (e.g. due to foliation or veins) can be detected as grain edges during the segmentation procedure. Therefore, large particles often appeared over-segmented. This resulted in the detection of several smaller particles instead of a single large particle. These over-segmentation errors were less frequent for small particles, whose detection was less influenced by rock texture details—apparently because of the limited image resolution. BASEGRAIN suffered from the same problem identified in PebbleCounts: it often did not merge shaded grain-surfaces into the object boundaries. This resulted in underestimation of the real size of grain features. Rock texture variations induced BASEGRAIN to detect fictitious edge patterns.

Non-detection of large particles might arise when BASEGRAIN was unable to reconstruct a closed grain boundary from these edge patterns. The large number of tunable parameters in BASEGRAIN makes the performance highly dependent on the operator's choices. Meticulous parameter tuning and visual checks of the quality of the object detection were performed in the present study, but it cannot be excluded that different operators could have obtained different results.

Systematic grain size underestimation from automated image-based methods induced by rock texture is an issue pinpointed by earlier studies (e.g., Strom et al., 2010). We found that PebbleCounts results were relatively less influenced by rock texture patterns than BASEGRAIN ones. Over-segmentation errors with BASEGRAIN may be even more severe when no parameter tuning is performed (see Chardon et al., 2022).

Among the routines considered here, GALET, based on deep-learning for object detection, emerged as the best-suited tool for grain size analysis when studying gravel bars from orthoimages. GALET was designed to fit the wide range of rock texture found in gravel-bed rivers and to conduct grain feature detection on long stream reaches. The object-detection performances of deep-learning methods (see Zhao et al., 2019), the recent implementation of deep learning in object-based grain-sizing techniques (e.g., Soloy et al., 2020; Chen et al., 2022) and the results of the present study indicate that the deep-learning technology may enable a step forward in automated optical granulometry.

4.3 Spatial variability of grain sizes

In Section 3.2, we provided quantitative information about the grain-size variability along the study reach. We conducted a statistical analysis on the d_{50} grid estimates and on the grain features detected by GALET. The accuracy evaluation of GALET revealed that this software performs at its best for grain size percentiles above d_{40}. Therefore, the d_{50} map presented in Figure 7 and the associated identified clusters should correspond closely to the real grain size variability on the field. Our analysis demonstrated how variable the bed-surface material size was in the surveyed floodplain. We found that grains were arranged in patches (i.e. spatial clusters), consistently with what has been observed for bars in gravel-bed rivers (e.g., Dietrich, 2006; Guerit et al., 2014).

5. CONCLUSIONS

We presented a benchmarking study of three object-based grain sizing models (BASEGRAIN, PebbleCountsAuto, and GALET) on a mountain river bed. The main difference between them was that GALET uses deep learning technology whereas the two others are based on image thresholding for grain segmentation. The three methods were applied to orthoimages obtained from UAV surveys and Structure-from-Motion photogrammetry. In-field estimates of grain sizes obtained using Fehr's line sampling technique served as a reference dataset to evaluate the accuracy of each grain sizing method. We supplemented the comparison by manually labelling grain features on the same orthoimages. By computing the grain size distributions within Fehr's (1987) framework, we ensured that all methods were comparable on the same footing.

Manual labelling provided estimates that were fully consistent with field measurements. BASEGRAIN and PebbleCountsAuto underestimated grain sizes on average, whereas GALET generally overestimated grain size percentiles.

We identified some limitations in the three models. BASEGRAIN often led to over-segmentation in grain features due to the rock-texture influence on object detection. PebbleCountsAuto outcomes were often affected by missed detections of large grains. Shaded grain faces and glacier flour also influenced grain detection in PebbleCountsAuto. The available image resolution prevented the detection of the smallest grain features with GALET.

Our case study confirms that automated grain-size measurement methods are cost-effective solutions for monitoring river-surface grain sizes with high spatial coverage. Such automated techniques are expected to progress in upcoming years with the advent of new grain-sizing procedures based on the deep learning technology.

We studied how grain sizes were distributed over the bed surface, by mapping characteristic grain sizes variations. Computing Moran’s I and LISA statistics revealed a significant grain-size clustering and the location of such grain-size patches (i.e. spatial clusters).. We believe that the combined use of automated grain sizing techniques (applied to large scale orthoimages) and spatial statistics will give a new impetus to understanding the processes that drive sediment sorting in gravel-bed rivers.

REFERENCES

Adams, J. (1979) 'Gravel Size Analysis from Photographs', *Journal of the Hydraulics Division*, 105(10), pp. 1247–1255. Available at: https://doi.org/10.1061/JYCEAJ.0005283.

Anselin, L. (1995) 'Local Indicators of Spatial Association—LISA', *Geographical Analysis*, 27(2), pp. 93–115. Available at: https://doi.org/10.1111/j.1538-4632.1995.tb00338.x.

Anselin, L. (1996) 'The Moran scatterplot as an ESDA tool to assess local instability in spatial association', in M. Fischer, H.J. Scholten, and D. Unwin (eds) *Spatial Analytical Perspectives on GIS*. 1st edn. Routledge. Available at: https://doi.org/10.1201/9780203739051.

Anselin, L., Syabri, I. and Kho, Y. (2010) 'GeoDa: An Introduction to Spatial Data Analysis', in M.M. Fischer and A. Getis (eds) *Handbook of Applied Spatial Analysis: Software Tools, Methods and Applications*. Berlin, Heidelberg: Springer, pp. 73–89. Available at: https://doi.org/10.1007/978-3-642-03647-7_5.

Auchincloss, A.H. *et al.* (2012) 'A Review of Spatial Methods in Epidemiology, 2000–2010', *Annual Review of Public Health*, 33(1), pp. 107–122. Available at: https://doi.org/10.1146/annurev-publhealth-031811-124655.

Baller, R.D. *et al.* (2001) 'Structural Covariates of U.s. County Homicide Rates: Incorporating Spatial Effects', *Criminology*, 39(3), pp. 561–588. Available at: https://doi.org/10.1111/j.1745-9125.2001.tb00933.x.

Bluck, B.J. (1979) 'Structure of coarse grained braided stream alluvium', *Earth and Environmental Science Transactions of The Royal Society of Edinburgh*, 70(10–12), pp. 181–221. Available at: https://doi.org/10.1017/S0080456800012795.

Bunte, K. and Abt, S.R. (2001) *Sampling Surface and Subsurface Particle-size Distributions in Wadable Gravel- and Cobble-bed Streams for Analyses in Sediment Transport, Hydraulics, and Streambed Monitoring*. U.S. Department of Agriculture, Forest Service, Rocky Mountain Research Station (General Techn. Report RMRS-GTR-74).

Buscombe, D. (2013) 'Transferable wavelet method for grain-size distribution from images of sediment surfaces and thin sections, and other natural granular patterns', *Sedimentology*, 60(7), pp. 1709–1732. Available at: https://doi.org/10.1111/sed.12049.

Buscombe, D. (2020) 'SediNet: a configurable deep learning model for mixed qualitative and quantitative optical granulometry', *Earth Surface Processes and Landforms*, 45(3), pp. 638–651. Available at: https://doi.org/10.1002/esp.4760.

Butler, J.B., Lane, S.N. and Chandler, J.H. (2001) 'Automated extraction of grain-size data from gravel surfaces using digital image processing', *Journal of Hydraulic Research*, 39(5), pp. 519–529. Available at: https://doi.org/10.1080/00221686.2001.9628276.

Carbonneau, P.E., Lane, S.N. and Bergeron, N.E. (2004) 'Catchment-scale mapping of surface grain size in gravel bed rivers using airborne digital imagery', *Water Resources Research*, 40(7). Available at: https://doi.org/10.1029/2003WR002759.

Carrivick, J.L. and Smith, M.W. (2019) 'Fluvial and aquatic applications of Structure from Motion photogrammetry and unmanned aerial vehicle/drone technology', *WIREs Water*, 6(1), p. e1328. Available at: https://doi.org/10.1002/wat2.1328.

Chardon, V., Piasny, G. and Schmitt, L. (2022) 'Comparison of software accuracy to estimate the bed grain size distribution from digital images: A test performed along the Rhine River', *River Research and Applications*, 38(2), pp. 358–367. Available at: https://doi.org/10.1002/rra.3910.

Chen, X., Hassan, M.A. and Fu, X. (2022) 'Convolutional neural networks for image-based sediment detection applied to a large terrestrial and airborne dataset', *Earth Surface Dynamics*, 10(2), pp. 349–366. Available at: https://doi.org/10.5194/esurf-10-349-2022.

Church, M., McLean, D.G. and Wolcott, J.F. (1987) 'River bed gravels: sampling and analysis.', in *Sediment transport in gravel-bed rivers*, pp. 43–88.

Daniels, M.D. and McCusker, M.H. (2010) 'Operator bias characterizing stream substrates using Wolman pebble counts with a standard measurement template', *Geomorphology*, 115(1), pp. 194–198. Available at: https://doi.org/10.1016/j.geomorph.2009.09.038.

Detert, M. and Weitbrecht, V. (2012) 'Automatic object detection to analyze the geometry of gravel grains - A free stand-alone tool', *River Flow 2012 - Proceedings of the International Conference on Fluvial Hydraulics*, 1, pp. 595–600.

Dietrich, W.E. *et al.* (2006) 'Sediment patches, sediment supply, and channel morphology', in G. Parker and M.H. Garcia (eds) *River Coastal and Estuarine Morphodynamics*, pp. 79–90. Available at: https://resolver.caltech.edu/CaltechAUTHORS:20130307-142351272.

Drăguţ, L. and Eisank, C. (2012) 'Automated object-based classification of topography from SRTM data', *Geomorphology*, 141–142, pp. 21–33. Available at: https://doi.org/10.1016/j.geomorph.2011.12.001.

Fehr, R. (1987) 'Geschiebeanalysen in Gebirgsflüssen: Umrechnung und vergleich von verschiedenen Analyseverfahren', *Mitteilungen der Versuchsanstalt fur Wasserbau, Hydrologie und Glaziologie an der Eidgenossischen Technischen Hochschule Zurich*, 92.

Ferguson, R. (2007) 'Flow resistance equations for gravel- and boulder-bed streams', *Water Resources Research*, 43(5). Available at: https://doi.org/10.1029/2006WR005422.

Garefalakis, P. *et al.* (2023) 'Comparison of three grain size measuring methods applied to coarse-grained gravel deposits', *Sedimentary Geology*, 446, p. 106340. Available at: https://doi.org/10.1016/j.sedgeo.2023.106340.

Getis, A. (2010) 'Spatial Autocorrelation', in M.M. Fischer and A. Getis (eds) *Handbook of Applied Spatial Analysis: Software Tools, Methods and Applications*, pp. 255–278. Available at: https://doi.org/10.1007/978-3-642-03647-7_14.

Graham, D.J. *et al.* (2010) 'Maximizing the accuracy of image-based surface sediment sampling techniques', *Water Resources Research*, 46(2). Available at: https://doi.org/10.1029/2008WR006940.

Graham, D.J., Rice, S.P. and Reid, I. (2005) 'A transferable method for the automated grain sizing of river gravels', *Water Resources Research*, 41(7). Available at: https://doi.org/10.1029/2004WR003868.

Guerit, L. *et al.* (2014) 'The Grain-size Patchiness of Braided Gravel-Bed Streams – example of the Urumqi River (northeast Tian Shan, China)', in *Advances in Geosciences*, pp. 27–39. Available at: https://doi.org/10.5194/adgeo-37-27-2014.

Hey, R.D. (1979) 'Flow Resistance in Gravel-Bed Rivers', *Journal of the Hydraulics Division*, 105(4), pp. 365–379. Available at: https://doi.org/10.1061/JYCEAJ.0005178.

Hodge, R., Richards, K. and Brasington, J. (2007) 'A physically-based bedload transport model developed for 3-D reach-scale cellular modelling', *Geomorphology*, 90(3), pp. 244–262. Available at: https://doi.org/10.1016/j.geomorph.2006.10.022.

Ibbeken, H. and Schleyer, R. (1986) 'Photo-sieving: A method for grain-size analysis of coarse-grained, unconsolidated bedding surfaces', *Earth Surface Processes and Landforms*, 11(1), pp. 59–77. Available at: https://doi.org/10.1002/esp.3290110108.

Kellerhals, R. and Bray, D.I. (1971) 'Sampling Procedures for Coarse Fluvial Sediments', *Journal of the Hydraulics Division*, 97(8), pp. 1165–1180. Available at: https://doi.org/10.1061/JYCEAJ.0003044.

Keulegan, G.H. (1938) 'Laws of turbulent flow in open channels', *Journal of Research of the National Bureau of Standards*, 21(6), pp. 707–741.

Knox, E.G. (1989) 'Detection of clusters', in P. Elliott (ed.) *Methodology of enquiries into disease clustering. London: Small Area Health Statistics Unit*, p. 20.

Konz, M. *et al.* (2011) 'Sediment transport modelling in a distributed physically based hydrological catchment model', *Hydrology and Earth*

System Sciences, 15(9), pp. 2821–2837. Available at: https://doi.org/10.5194/hess-15-2821-2011.

Lang, N. *et al.* (2021) ‘GRAINet: mapping grain size distributions in river beds from UAV images with convolutional neural networks’, *Hydrology and Earth System Sciences*, 25(5), pp. 2567–2597. Available at: https://doi.org/10.5194/hess-25-2567-2021.

Lisle, T.E. and Madej, M.A. (1992) ‘Spatial variation in armouring in a channel with high sediment supply’, *Dynamics of Gravel-bed Rivers.*, pp. 277–293.

Maimaitijiang, M. *et al.* (2020) ‘Soybean yield prediction from UAV using multimodal data fusion and deep learning’, *Remote Sensing of Environment*, 237, p. 111599. Available at: https://doi.org/10.1016/j.rse.2019.111599.

Mair, D. *et al.* (2022) ‘Grain size of fluvial gravel bars from close-range UAV imagery – uncertainty in segmentation-based data’, *Earth Surface Dynamics*, 10(5), pp. 953–973. Available at: https://doi.org/10.5194/esurf-10-953-2022.

McEwan, I.K. *et al.* (2000) ‘Estimating the size composition of sediment surfaces through image analysis’, *Proceedings of the Institution of Civil Engineers - Water and Maritime Engineering*, 142(4), pp. 189–195. Available at: https://doi.org/10.1680/wame.2000.142.4.189.

Moran, P.A.P. (1950) ‘Notes on Continuous Stochastic Phenomena’, *Biometrika*, 37(1/2), pp. 17–23. Available at: https://doi.org/10.2307/2332142.

Mörtl, C. *et al.* (2022) ‘GALET: A deep learning image segmentation model for drone-based grain size analysis of gravel bars’, in *Proceedings of the 39th IAHR World Congress—From Snow To Sea*. Available at: https://doi.org/10.3850/IAHR-39WC252171192022895.

Piégay, H. *et al.* (2020) ‘Remotely sensed rivers in the Anthropocene: state of the art and prospects’, *Earth Surface Processes and Landforms*, 45(1), pp. 157–188. Available at: https://doi.org/10.1002/esp.4787.

Purinton, B. and Bookhagen, B. (2019) ‘Introducing *PebbleCounts*: a grain-sizing tool for photo surveys of dynamic gravel-bed rivers’, *Earth Surface Dynamics*, 7(3), pp. 859–877. Available at: https://doi.org/10.5194/esurf-7-859-2019.

Purinton, B. and Bookhagen, B. (2021) ‘Tracking Downstream Variability in Large Grain-Size Distributions in the South-Central Andes’, *Journal of Geophysical Research: Earth Surface*, 126(8), p. e2021JF006260. Available at: https://doi.org/10.1029/2021JF006260.

Ramirez, J.A. *et al.* (2022) ‘Modelling the long-term geomorphic response to check dam failures in an alpine channel with CAESAR-Lisflood’, *International Journal of Sediment Research*, 37(5), pp. 687–700. Available at: https://doi.org/10.1016/j.ijsrc.2022.04.005.

Recking, A. (2013) 'Simple Method for Calculating Reach-Averaged Bed-Load Transport', *Journal of Hydraulic Engineering*, 139(1), pp. 70–75. Available at: https://doi.org/10.1061/(ASCE)HY.1943-7900.0000653.

Recking, A. (2016) 'A generalized threshold model for computing bed load grain size distribution', *Water Resources Research*, 52(12), pp. 9274–9289. Available at: https://doi.org/10.1002/2016WR018735.

Rice, S. and Church, M. (1996) 'Sampling surficial fluvial gravels; the precision of size distribution percentile sediments', *Journal of Sedimentary Research*, 66(3), pp. 654–665. Available at: https://doi.org/10.2110/jsr.66.654.

Rickenmann, D. (1997) 'Sediment transport in Swiss torrents', *Earth Surface Processes and Landforms*, 22(10), pp. 937–951. Available at: https://doi.org/10.1002/(SICI)1096-9837(199710)22:10<937::AID-ESP786>3.0.CO;2-R.

Rickenmann, D. and McArdell, B.W. (2007) 'Continuous measurement of sediment transport in the Erlenbach stream using piezoelectric bedload impact sensors', *Earth Surface Processes and Landforms*, 32(9), pp. 1362–1378. Available at: https://doi.org/10.1002/esp.1478.

Rickenmann, D. and Recking, A. (2011) 'Evaluation of flow resistance in gravel-bed rivers through a large field data set', *Water Resources Research*, 47(7). Available at: https://doi.org/10.1029/2010WR009793.

Schlunegger, F., Delunel, R. and Garefalakis, P. (2020) 'Short communication: Field data reveal that the transport probability of clasts in Peruvian and Swiss streams mainly depends on the sorting of the grains', *Earth Surface Dynamics*, 8(3), pp. 717–728. Available at: https://doi.org/10.5194/esurf-8-717-2020.

Schneider, J.M. *et al.* (2016) 'Bed load transport in a very steep mountain stream (Riedbach, Switzerland): Measurement and prediction', *Water Resources Research*, 52(12), pp. 9522–9541. Available at: https://doi.org/10.1002/2016WR019308.

Smart, G.M. and Jaeggi, M.N.R. (1983) 'Sediment transport on steep slopes', *Mitteilungen der Versuchsanstalt für Wasserbau, Hydrologie und Glaziologie*, 64, pp. 91–191.

Soloy, A. *et al.* (2020) 'A Deep Learning-Based Method for Quantifying and Mapping the Grain Size on Pebble Beaches', *Remote Sensing*, 12(21), p. 3659. Available at: https://doi.org/10.3390/rs12213659.

Strickler, A. (1923) 'Beiträge zur Frage der Geschwindigkeitsformel und der Rauhigkeitszahlen für Ströme, Kanäle und geschlossene Leitungen', *Eidg. Amt für Wasserwirtsch.*

Strom, K.B., Kuhns, R.D. and Lucas, H.J. (2010) ‘Comparison of Automated Image-Based Grain Sizing to Standard Pebble-Count Methods’, *Journal of Hydraulic Engineering*, 136(8), pp. 461–473. Available at: https://doi.org/10.1061/(ASCE)HY.1943-7900.0000198.

Travaglini, E. *et al.* (2015) ‘Analysis of Sediment Transport from Recorded Signals of Sediments in a Gravel-Bed River: Role of Sediment Availability’, in G. Lollino et al. (eds) *Engineering Geology for Society and Territory - Volume 3*. Cham: Springer International Publishing, pp. 477–481. Available at: https://doi.org/10.1007/978-3-319-09054-2_98.

Warrick, J.A. *et al.* (2009) ‘Cobble cam: grain-size measurements of sand to boulder from digital photographs and autocorrelation analyses’, *Earth Surface Processes and Landforms*, 34(13), pp. 1811–1821. Available at: https://doi.org/10.1002/esp.1877.

Westoby, M.J. *et al.* (2012) ‘“Structure-from-Motion” photogrammetry: A low-cost, effective tool for geoscience applications’, *Geomorphology*, 179, pp. 300–314. Available at: https://doi.org/10.1016/j.geomorph.2012.08.021.

Wolman, G. (1954) ‘A method of sampling coarse river-bed material’, *Eos, Transactions American Geophysical Union*, 35(6), pp. 951–956. Available at: https://doi.org/10.1029/TR035i006p00951.

Woodget, A.S. and Austrums, R. (2017) ‘Subaerial gravel size measurement using topographic data derived from a UAV-SfM approach’, *Earth Surface Processes and Landforms*, 42(9), pp. 1434–1443. Available at: https://doi.org/10.1002/esp.4139.

Woodget, A.S., Fyffe, C. and Carbonneau, P.E. (2018) ‘From manned to unmanned aircraft: Adapting airborne particle size mapping methodologies to the characteristics of sUAS and SfM’, *Earth Surface Processes and Landforms*, 43(4), pp. 857–870. Available at: https://doi.org/10.1002/esp.4285.

Zhao, Z.-Q. *et al.* (2019) ‘Object Detection With Deep Learning: A Review’, *IEEE Transactions on Neural Networks and Learning Systems*, 30(11), pp. 3212–3232. Available at: https://doi.org/10.1109/TNNLS.2018.2876865.