M_{split} estimation as a method for processing heterogeneous data

Patrycja WYSZKOWSKA, Robert DUCHNOWSKI, Poland

Key words: M_{split} estimation, heterogeneous data, laser scanning, outliers

SUMMARY

M_{split} estimation is a modern estimation method that is a development of the maximum likelihood estimation. The basic assumption of M_{split} estimation is that an observation set is a mixture of realizations of at least two different random variables. In other words, the observation set might consist of different observation groups (aggregations), which differ from each other in location parameters. The main objective of M_{split} estimation is to assess such parameters as different versions of the split functional model parameters. The first and basic variant of M_{split} estimation is called the squared M_{split} estimation, and it can be derived from the assumption that the measurement errors are normally distributed. Since this variant is sensitive to outlying observations, the absolute M_{split} estimation has been introduced. This variant can be regarded as the least absolute deviation method development. It can be proved that the absolute M_{split} estimation is less sensitive to outlying observations than the squared M_{split} estimation. Both variants found several practical applications in geodetic data processing, e.g., deformation analysis, detection of gross errors, coordinates transformation, or laser scanning data processing. The last application seems especially interesting nowadays when the LiDAR technique becomes very popular. The laser scanning results, usually in the form of a point cloud, often contain measurements of different objects, e.g., terrain surface, buildings, engineering structures, or vegetation cover. Therefore, point clouds should be considered heterogeneous observation sets. Thus, such sets seem adequate to be processed by applying M_{split} estimation. The paper shows the practical application of M_{split} estimation in processing laser scanning data; approximation of one surface or two surfaces from a single observation set. The results are compared to the least squares estimation. One can conclude that both variants of M_{split} estimation might provide better results, and for some types of point clouds, they should be recommended.

M_{split} estimation as a method for processing heterogeneous data

Patrycja WYSZKOWSKA, Robert DUCHNOWSKI, Poland

1. INTRODUCTION

Modern measurement techniques often provide observation data that consist of thousands or even millions of measured points. Such data might result from the application of LiDAR (light detection and ranging) systems. Measurement data usually contain observations concerning different objects, e.g., terrain surface, buildings, engineering structures, or vegetation cover; hence data are heterogeneous. Processing heterogeneous data might cause problems if we do not realize that only part of the measured points concerns the study object. The question arises, how to separate the data from the object from disturbing data. One can apply one of the data cleaning procedures; however, they might sometimes fail (Baselga 2011; Chen et al. 2017). On the other hand, heterogeneous sets seem adequate to be processed by applying M_{split} estimation, a modern estimation method that develops the maximum likelihood estimation (Wiśniewski 2009). The basic assumption of M_{split} estimation is that an observation set is a mixture of realizations of at least two different random variables. In other words, the observation set might consist of varying observation groups (aggregations), which differ from each other in location parameters. The main objective of M_{split} estimation is to assess such parameters as different versions of the split functional model parameters. The first and basic variant of M_{split} estimation is called the squared M_{split} estimation (SMS estimation). It can be derived from the assumption that the measurement errors are normally distributed (Wiśniewski 2009). Since this variant is sensitive to outlying observations, the absolute M_{split} estimation (AMS estimation) has been introduced (Wyszkowska and Duchnowski 2019). Considering the form of the objective function, that variant can be regarded as the least absolute deviation method development. It can be proved that the absolute M_{split} estimation is less sensitive to outlying observations than the squared M_{split} estimation (Wyszkowska and Duchnowski 2019; Wyszkowska et al. 2021). Both variants found several practical applications in geodetic data processing, e.g., deformation analysis (e.g., Wiśniewski 2009; Wiśniewski and Zienkiewicz 2016; Zienkiewicz et al. 2017; Wyszkowska and Duchnowski 2019), detection of gross errors (Li et al. 2013), circle object detection (Janowski 2018; Baselga et al. 2021), laser scanning data (Janowski and Rapiński 2013; Błaszczak-Bąk et al. 2015; Janicka et al. 2020; Wyszkowska et al. 2021), processing coordinates transformation (Janicka and Rapiński 2013), S-transformation (Nowel 2019; Guo et al. 2020), linear regression (Wiśniewski 2010; Wyszkowska and Duchnowski 2019), or marine navigation (Zienkiewicz and Czaplewski 2017; Czaplewski et al. 2019).

This paper focuses on processing heterogeneous data, namely point clouds obtained from LiDAR systems, by applying M_{split} estimation. We assume that data contain some disturbing points resulting from measurements of different objects, not the study object. In that context, such mismeasured points might be classified as outliers (Carrilho and Galo 2019; Wyszkowska et al. 2021). When applying M_{split} estimation, one can distinguish two main variants. The natural approach to that estimation method is estimating two (or more) parameter variants describing two surfaces, profiles, etc. In the latter approach, one is interested only in one solution, and the

second solution is out of interest as describing the location or placement of outliers. Those two approaches are addressed by simulating different heterogeneous observation sets and estimating the functional model parameters by applying the methods in question. Finally, the results are compared to outcomes of more conventional variants of M-estimation, including least squares estimation (LS estimation).

2. THEORETICAL FOUNDATIONS

Considering geodetic observations, one usually uses their functional model in the following linear form

$$\mathbf{y} = \mathbf{A}\mathbf{X} + \mathbf{v} \tag{1}$$

where: $\mathbf{y} = \left[y_1, ..., y_n\right]^T$ is an observation vector, $\mathbf{v} = \left[v_1, ..., v_n\right]^T$ is a vector of random errors, $\mathbf{X} = \left[X_1, ..., X_m\right]^T$ is a parameter vector, **A** is a known rectangular matrix of size $n \times m$; here, it is assumed to be full column rank. Here it holds that the expected value $E(\mathbf{y}) = \mathbf{A}\mathbf{X}$. If one accepts the same accuracy for all observations, their weight matrix equals the identity matrix $(\mathbf{P} = \mathbf{I})$. In such a case, the least squares (LS) estimate of the parameter vector can be given as follows

$$\hat{\mathbf{X}}_{LS} = \left(\mathbf{A}^T\mathbf{A}\right)^{-1}\mathbf{A}^T\mathbf{y} \tag{2}$$

The model of Eq. (1) is not generally applicable in the case of M_{split} estimation. Considering the general assumption of that method, the traditional functional model should be split into two competitive models (Wiśniewski 2009)

$$\mathbf{y} = \mathbf{A}\mathbf{X} + \mathbf{v} \;\Rightarrow\; \begin{cases} \mathbf{y} = \mathbf{A}\mathbf{X}_{(1)} + \mathbf{v}_{(1)} \\ \mathbf{y} = \mathbf{A}\mathbf{X}_{(2)} + \mathbf{v}_{(2)} \end{cases} \tag{3}$$

where: $\mathbf{X}_{(1)}$ and $\mathbf{X}_{(2)}$ are the competitive versions of the parameter vector **X**, $\mathbf{v}_{(1)}$ and $\mathbf{v}_{(2)}$ are the competitive versions of the vector of random errors **v**. During the estimation process, each observation is assigned to either of the split functional models. Here we assume two such models; however, it is also possible to split the conventional model into *q* competitive ones $(q < n)$, which leads to $M_{split(q)}$ estimation (Wiśniewski 2010).

The competitive versions of parameters are computed in an iterative process that is based on the Newton-Raphson method. One can apply either of two schemes (Wyszkowska and Duchnowski 2020):

- traditional iterative process

$$\begin{aligned} \mathbf{X}_{(1)}^{j} &= \mathbf{X}_{(1)}^{j-1} + d\mathbf{X}_{(1)}^{j} = \mathbf{X}_{(1)}^{j-1} - \left[\mathbf{H}_{(1)}\left(\mathbf{X}_{(1)}^{j-1}, \mathbf{X}_{(2)}^{j-1}\right)\right]^{-1}\mathbf{g}_{(1)}\left(\mathbf{X}_{(1)}^{j-1}, \mathbf{X}_{(2)}^{j-1}\right) \\ \mathbf{X}_{(2)}^{j} &= \mathbf{X}_{(2)}^{j-1} + d\mathbf{X}_{(2)}^{j} = \mathbf{X}_{(2)}^{j-1} - \left[\mathbf{H}_{(2)}\left(\mathbf{X}_{(1)}^{j-1}, \mathbf{X}_{(2)}^{j-1}\right)\right]^{-1}\mathbf{g}_{(2)}\left(\mathbf{X}_{(1)}^{j-1}, \mathbf{X}_{(2)}^{j-1}\right) \end{aligned} \tag{4}$$

- parallel iterative process

$$\begin{aligned}\mathbf{X}_{(1)}^{j} &= \mathbf{X}_{(1)}^{j-1} + d\mathbf{X}_{(1)}^{j} = \mathbf{X}_{(1)}^{j-1} - \left[\mathbf{H}_{(1)}\left(\mathbf{X}_{(1)}^{j-1}, \mathbf{X}_{(2)}^{j-1}\right)\right]^{-1} \mathbf{g}_{(1)}\left(\mathbf{X}_{(1)}^{j-1}, \mathbf{X}_{(2)}^{j-1}\right) \\ \mathbf{X}_{(2)}^{j} &= \mathbf{X}_{(2)}^{j-1} + d\mathbf{X}_{(2)}^{j} = \mathbf{X}_{(2)}^{j-1} - \left[\mathbf{H}_{(2)}\left(\mathbf{X}_{(1)}^{j-1}, \mathbf{X}_{(2)}^{j-1}\right)\right]^{-1} \mathbf{g}_{(2)}\left(\mathbf{X}_{(1)}^{j-1}, \mathbf{X}_{(2)}^{j-1}\right)\end{aligned} \tag{5}$$

where: $d\mathbf{X}_{(l)}$ is an increment to parameter vector, $\mathbf{H}_{(l)}\left(\mathbf{X}_{(1)}, \mathbf{X}_{(2)}\right)$ are the Hessians, $\mathbf{g}_{(l)}\left(\mathbf{X}_{(1)}, \mathbf{X}_{(2)}\right)$ are the gradients, l is equal to 1 or 2. In both schemes, the Hessians and gradients are computed in the same following way

$$\begin{aligned}\mathbf{H}_{(1)}\left(\mathbf{X}_{(1)}, \mathbf{X}_{(2)}\right) &= 2\mathbf{A}^{T}\mathbf{w}_{(1)}\left(\mathbf{v}_{(1)}, \mathbf{v}_{(2)}\right)\mathbf{A} \\ \mathbf{H}_{(2)}\left(\mathbf{X}_{(1)}, \mathbf{X}_{(2)}\right) &= 2\mathbf{A}^{T}\mathbf{w}_{(2)}\left(\mathbf{v}_{(1)}, \mathbf{v}_{(2)}\right)\mathbf{A}\end{aligned} \tag{6}$$

$$\begin{aligned}\mathbf{g}_{(1)}\left(\mathbf{X}_{(1)}, \mathbf{X}_{(2)}\right) &= -2\mathbf{A}^{T}\mathbf{w}_{(1)}\left(\mathbf{v}_{(1)}, \mathbf{v}_{(2)}\right)\mathbf{v}_{(1)} \\ \mathbf{g}_{(2)}\left(\mathbf{X}_{(1)}, \mathbf{X}_{(2)}\right) &= -2\mathbf{A}^{T}\mathbf{w}_{(2)}\left(\mathbf{v}_{(1)}, \mathbf{v}_{(2)}\right)\mathbf{v}_{(2)}\end{aligned} \tag{7}$$

The matrices $\mathbf{w}_{(l)}\left(\mathbf{v}_{(1)}, \mathbf{v}_{(2)}\right)$ are computed in the subsequent iterative steps by applying the weight functions related to the variant of M_{split} estimates

$$\begin{aligned}\mathbf{w}_{(1)}\left(\mathbf{v}_{(1)}, \mathbf{v}_{(2)}\right) &= \mathrm{diag}\left[w_{(1)}\left(v_{1(1)}, v_{1(2)}\right), ..., w_{(1)}\left(v_{n(1)}, v_{n(2)}\right)\right] \\ \mathbf{w}_{(2)}\left(\mathbf{v}_{(1)}, \mathbf{v}_{(2)}\right) &= \mathrm{diag}\left[w_{(2)}\left(v_{1(1)}, v_{1(2)}\right), ..., w_{(2)}\left(v_{n(1)}, v_{n(2)}\right)\right]\end{aligned} \tag{8}$$

where: $\mathrm{diag}(\circ)$ is a diagonal matrix. The traditional process is dedicated to the variants of M_{split} estimation that use the mutual cross weighting, namely to the variants for which weight function $w_{(1)}\left(v_{i(1)}, v_{i(2)}\right)$ depends only on $v_{i(2)}$, and $w_{(2)}\left(v_{i(1)}, v_{i(2)}\right)$ only on $v_{i(2)}$. If the weight functions are defined in other ways, then one should apply the parallel iterative process (Wyszkowska and Duchnowski 2019, 2020).

So far, two main variants of M_{split} estimation have been introduced. The first and primary method in that context is called the squared M_{split} estimation (SMS estimation), for which the weight functions are written as follows (Wiśniewski 2009)

$$\begin{cases} w_{(1)}\left(v_{i(1)}, v_{i(2)}\right) = v_{i(2)}^{2} \\ w_{(2)}\left(v_{i(1)}, v_{i(2)}\right) = v_{i(1)}^{2} \end{cases} \tag{9}$$

The second variant, the absolute M_{split} estimation (AMS estimation), was meant to be less sensitive to outlying observations than SMS estimation. Its objective function is derived from L_1 norm condition, and the weight functions of AMS estimation are derived in the following forms (Wyszkowska and Duchnowski 2019)

$$
w_{(1)}\left(v_{i(1)}, v_{i(2)}\right)=\begin{cases}\dfrac{\left|v_{i(2)}\right|}{2c} & \text{for } \left|v_{i(1)}\right|<c \\ \dfrac{\left|v_{i(2)}\right|}{2\left|v_{i(1)}\right|} & \text{for } \left|v_{i(1)}\right|\geq c\end{cases}
$$

$$
w_{(2)}\left(v_{i(1)}, v_{i(2)}\right)=\begin{cases}\dfrac{\left|v_{i(1)}\right|}{2c} & \text{for } \left|v_{i(2)}\right|<c \\ \dfrac{\left|v_{i(1)}\right|}{2\left|v_{i(2)}\right|} & \text{for } \left|v_{i(2)}\right|\geq c\end{cases} \tag{10}
$$

where: c is an assumed small positive constant.

The differences in the forms of the weight functions of both M_{split} estimation variants are significant. They not only imply the possible scheme of getting a solution but also influence the main properties of those variants, which will be addressed in the following sections of the paper.

3. EMPIRICAL TESTS

Let us compare two variants of M_{split} estimation and LS estimation in processing LiDAR data which are heterogeneous. We consider two main issues. The first one concerns the situation when an observation set of airborne laser scanning (ALS) data is a mixture of measured points at two different surfaces, and one is interested in estimating both surfaces. The second issue addresses the problem of outlying observations in terrestrial laser scanning (TLS) data.

In the first numerical example, we consider simulated data that refer to two surfaces obtained by the ALS technique. Such observations (usually one point cloud of combined point clouds) might result from, for example, measuring the terrain (denoted *P*) covered by the vegetation (denoted *P'*). Let us assume that the profiles might describe the declared surfaces well enough. Hence, let us select 100 measurements that concern one example profile of 50 m long. For the simulation, we assume that the terrain profile is described by the second-degree polynomial (the assumed coefficients $c_2 = 0.002$, $c_1 = -0.07$, $c_0 = 1$), and the vegetation cover by the second-degree polynomial of the coefficients $c_2 = 0.001$, $c_1 = 0$, $c_0 = 1$. The observations are considered normally distributed, and their errors have the expected value of 0 mm and the standard deviation of 50 mm, which are acceptable for ALS data (Crespo-Peremarch et al. 2018). It is also important to realize that some measured points might not concern the declared surfaces but other objects in the study area. Thus, they should be regarded as outliers in the context of estimating the terrain or vegetation profiles. For that reason, simulated data might also be randomly disturbed by gross errors from 0.20 m to 0.80 m. Taking the mentioned source of outliers, one assumes that gross errors concerning the terrain profile are considered positive. If gross errors involve the vegetation cover, they lead to negative outliers (Carrilho and Galo 2019). The observation sets are simulated in three variants: A – observation set free of outliers, B – 10% of outliers within the observation set, and C – 30% of such observations.

The simulated observation sets can be processed by M_{split} estimation. However, when applying the classical methods, one should consider two groups separately – observations of the terrain profile and observations of the vegetation cover. The problem is that we do not know the actual data division (the real assignment of each observation to either of the subsets). Thus, the subsets should be created a little bit artificially: consider ten intervals of 5 m long, in each interval select the largest-valued observations, accounting the half of all. They are assigned as measurements of the vegetation cover (denoted as $\mathbf{y}_2$), whereas the other observations as measurements of the terrain profile (denoted as $\mathbf{y}_1$). The simulated observation sets and their division into the subsets are presented in Fig. 1.

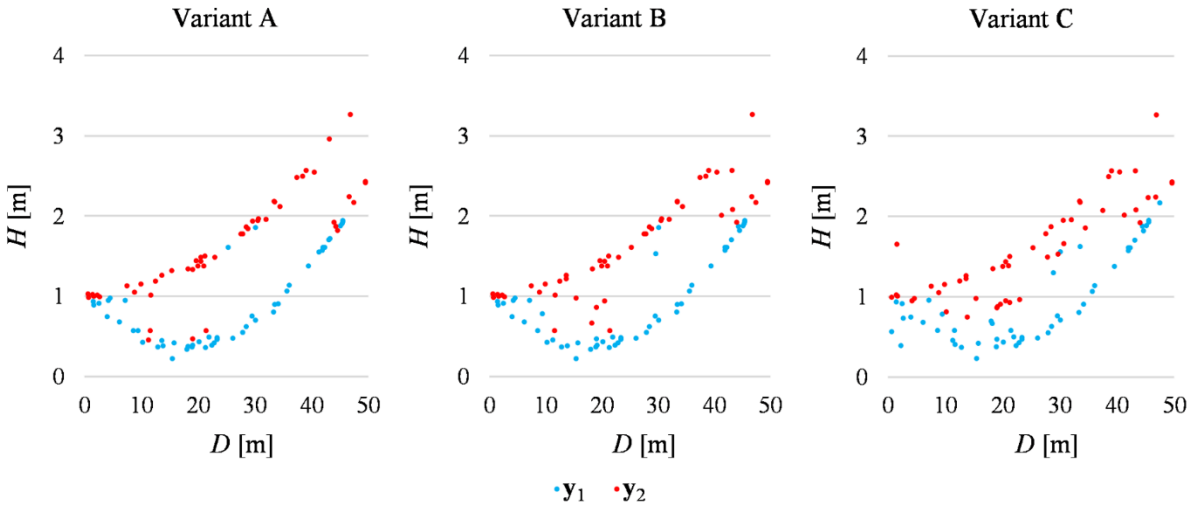

Fig. 1. Simulated observation sets in Variants A, B, and C

The simulated ALS data, the whole sets in the case of M_{split} estimation and the subsets $\mathbf{y}_1$ and $\mathbf{y}_2$ in LS estimation, are the base for estimating the polynomial coefficients that describe both profiles – the terrain and the vegetation cover. Such estimated coefficients allow one to compute the estimated profiles presented in Fig. 2. When comparing the estimated profiles with the respective simulated ones, the best-fitted profiles are those obtained by applying AMS estimation. SMS estimation cannot deal with outlying observations. It provides good results only in Variant A. LS estimation fails in all variants, and its effects are far from the simulated profiles. To describe the results in a better way, one can compute the accuracy of the fit of the estimated profiles to the simulated ones. Thus, the root-mean-square deviation (*RMSD*) be calculated by applying the following formula (e.g., Wyszkowska et al. 2021)

$$RMSD\left(\hat{H}\right)=\sqrt{\frac{\sum_{i=0}^{n}\left(\hat{H}_i-H_i\right)^2}{n}} \tag{11}$$

where: $\hat{H}_i$ are the estimated heights, and H_i are the simulated heights. Here n = 501, which is the number of points for which such heights are calculated for distances $D_j = j \cdot 0.1$ m

$(j = 0, ..., 500)$. *RMSD*s determined for all estimated profiles and variants are presented in Table 1, where $\hat{H}_{(1)}$ concerns the heights of the estimated terrain profiles whereas $\hat{H}_{(2)}$ the heights of the estimated profile of the vegetation cover. The obtained values confirm the conclusions resulting from the simple graphical analysis of profiles in Fig. 2. LS estimates are much more inaccurate than M_{split} estimates. Table 1 also shows that in Variant A the best-fitted profiles are obtained by applying SMS estimation.

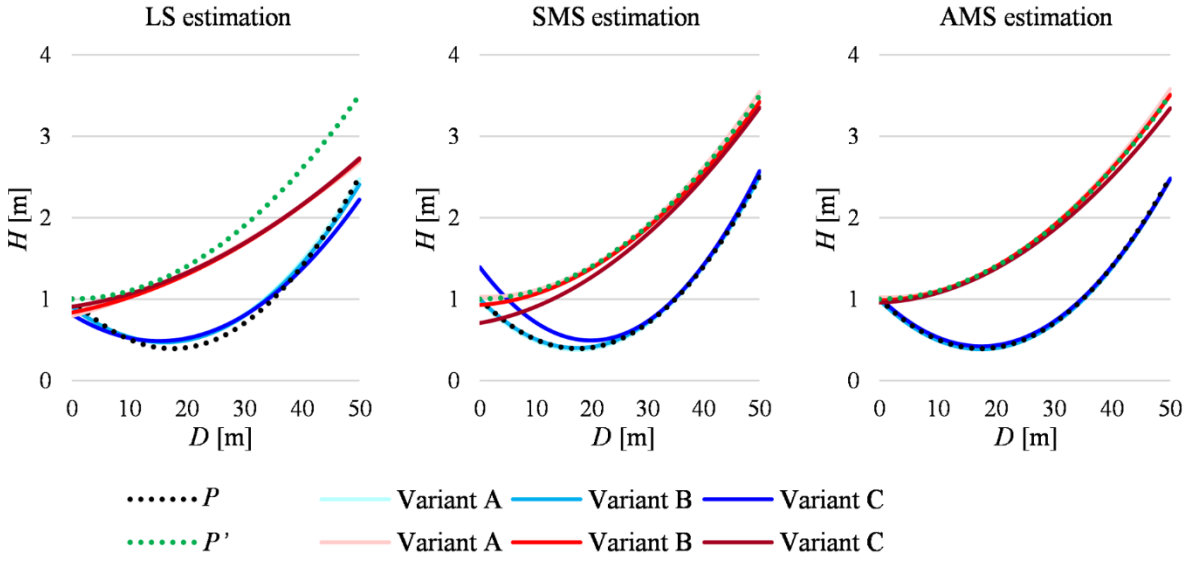

Fig. 2. Simulated and estimated profiles of terrain (*P*) and vegetation cover (*P*') in Variants A, B, and C

Table 1. Accuracy of fit of the profiles in Variants A, B, and C

Method	Parameter	Variant A	Variant B	Variant C
LS estimation	$RMSD\left(\hat{H}_{(1)}\right)$ [m]	0.074	0.069	0.109
	$RMSD\left(\hat{H}_{(2)}\right)$ [m]	0.333	0.329	0.317
SMS estimation	$RMSD\left(\hat{H}_{(1)}\right)$ [m]	0.009	0.012	0.155
	$RMSD\left(\hat{H}_{(2)}\right)$ [m]	0.018	0.042	0.156
AMS estimation	$RMSD\left(\hat{H}_{(1)}\right)$ [m]	0.010	0.010	0.022
	$RMSD\left(\hat{H}_{(2)}\right)$ [m]	0.029	0.007	0.070

The application of LS estimation to processing the subsets in that example might be questionable. Fig. 1 shows that the declared subsets contain outlying observations that do not result from the gross error occurrence in all variants. Some terrain observations are included in

the subsets of the vegetation cover observations and *vice versa*. Thus, let the subsets be processed by applying robust estimation methods, namely two variants of M-estimation, the Huber or Tukey method with the steering parameters k = 2 and 6, respectively (e.g., Baselga 2007; Ge et al. 2013). Table 2 presents the accuracy of the fit of the estimated profiles based on M-estimates of the respective polynomials' coefficients. Application of the robust M-estimates often improves the accuracy of the fit of the estimated terrain profile (in relation to LS estimation); however, the improvement is negligible. As for the estimated vegetation cover profiles, the fit accuracy becomes even worse. The results of the robust M-estimation are much worse than the results of M_{split} estimation. One can say that in the case of the presented heterogeneous data, M_{split} estimation overperforms not only LS estimation but also conventional robust solutions.

Table 2. Accuracy of fit of the profiles determined by M-estimation in Variants A, B, and C

Method	Parameter	Variant A	Variant B	Variant C
Huber	$RMSD\left(\hat{H}_{(1)}\right)$ [m]	0.038	0.041	0.100
	$RMSD\left(\hat{H}_{(2)}\right)$ [m]	0.338	0.335	0.341
Tukey	$RMSD\left(\hat{H}_{(1)}\right)$ [m]	0.045	0.046	0.102
	$RMSD\left(\hat{H}_{(2)}\right)$ [m]	0.331	0.330	0.333

The second example considered here concerns the determination of a beam deformation by applying the TLS technique. Such an approach is very useful in many practical problems, such as when steel beams are parts of roof constructions and are often out of the reach of conventional measurement methods (Gordon and Lichti 2007; Cabaleiro et al. 2015). Here, we consider the case study presented in the latter paper. Thus, let the steel beam of 5870 mm length be deformed under the asymmetric load of 4 kN (at 1940 mm) and 2 kN (at 3810 mm). The beam deformation was measured by applying the contact instruments at five chosen points (located at the center of the beam flange), resulting in the following beam deflections: 0.9 mm (at 100 mm), 21.5 mm (at 1940 mm), 21.9 mm (at 2935 mm), 19.2 mm (at 3810 mm), 1.0 mm (at 5770 mm). Such values are a base for computing the reference polynomial (the fourth-degree polynomial model that reflects the actual beam deflection). Such a polynomial can be estimated using simulated TLS data by assuming that the accuracy of measurements is 2 mm and the resolution 10 mm. Thus, one considers 587 points placed randomly on the beam surface with random errors normally distributed with the assumed standard deviation. One can find the full description of the experiment in (Cabaleiro et al. 2015).

The TLS measurements of a beam would result in a homogeneous point cloud. Thus, such data should not be processed by using M_{split} estimation. Let us now assume that some observations are outlying. Here, one can consider two possible sources of outliers: measurements of construction elements different from the beam under investigation and measurements of contaminations of the beam surface (like dirt, peeling paints, corrosion). Thus, we consider three variants of the simulated observation sets: Variant A and B – contain one or

two groups of 50 ”strange” observations, respectively (the ”strange” observations can be regarded as outliers of high magnitudes), Variant C – 20% of the beam observations are additionally contaminated by gross errors from the interval $\langle 3\,\mathrm{mm}, 10\,\mathrm{mm}\rangle$ (simulating the measurements of the beam contamination – outliers of low or moderate magnitude). The simulated observation sets are shown in Fig. 3.

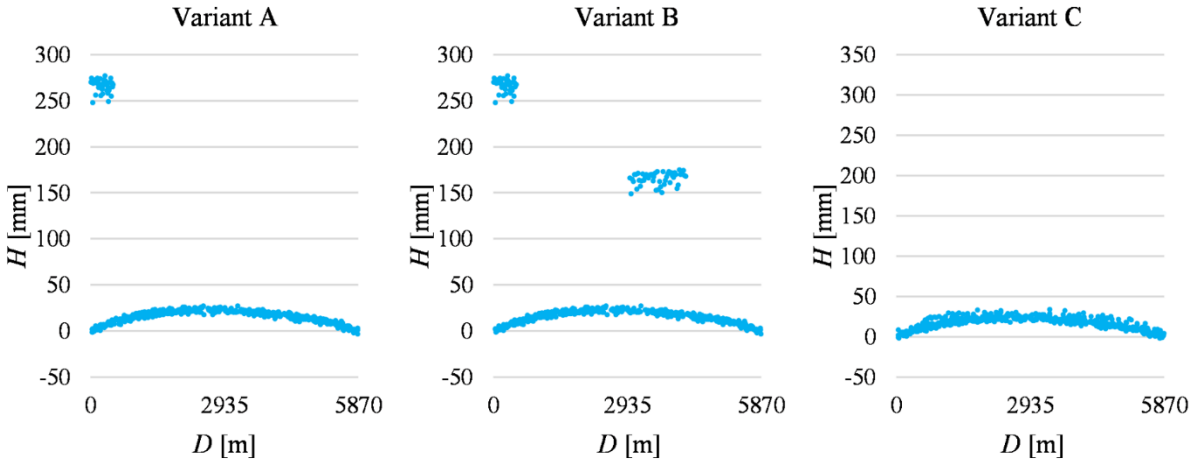

Fig. 3. Simulated observation sets in Variants A, B, and C

Let now the beam deflection be modeled by applying fourth-degree polynomials, which stems from the theoretical consideration of beam deflection under the load (e.g., Gordon and Lichti 2007; Holst et al. 2014). Let the polynomial coefficients be estimated by applying both variants of M_{split} estimation and LS estimation just for comparison. The main difference between that example and the previous one is that we now are interested in estimating only one surface (the beam flange). It means that in the case of M_{split} estimation, one should consider only one solution; the second one is out of interest as describing the location of outlying observations. The choice of the correct solution is rather apparent; one should take the polynomial in which the graph is located lower.

The estimation results, namely the polynomial models of the beam deformation, are presented in Fig. 4. M_{split} estimation variants overperform LS estimation in Variants A and B, which means that they can both cope with high-magnitude outliers. One can also notice that AMS estimation provides significantly better results than SMS estimation in Variant B. In Variant C, the model obtained by applying SMS estimation seems worse, contrary to AMS estimation, which still provides a good solution. SMS estimation has “problems” with modeling the beam deflection in its first half. It probably stems from the location of multiple outliers in that data part and the method's sensitivity to such an observation type. It is not surprising that LS estimation cannot deal with outliers of a high magnitude; hence its results in Variants A and B are far from the reference model for both polynomial variants. In Variant C, the result is much better and very close to the reference model; however, it is worse than the model obtained from AMS estimation.

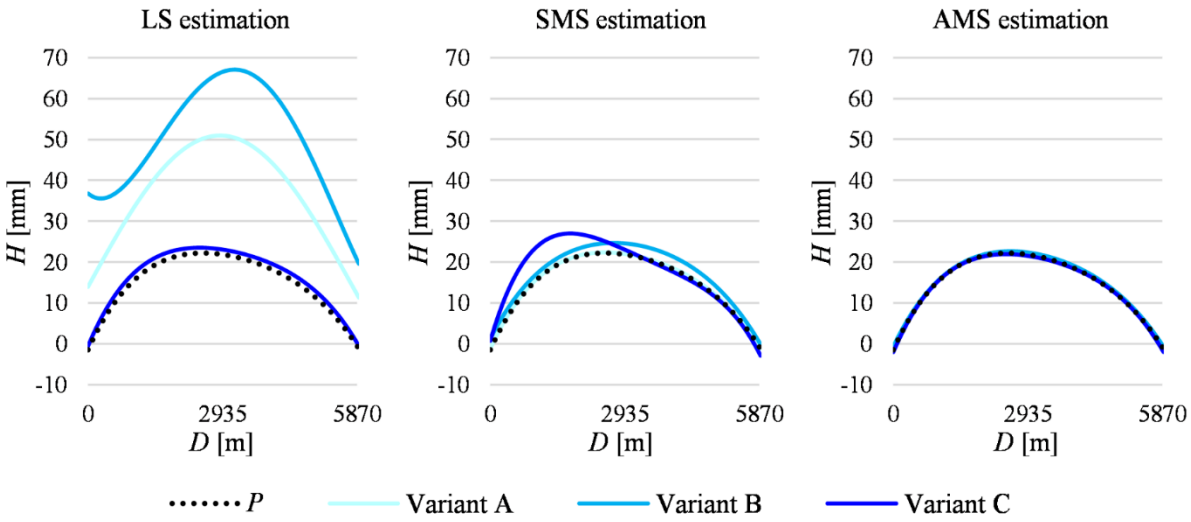

Fig. 4. Simulated profile P and estimated ones in Variants A, B, and C

To compare the fit of all estimated models in the reference one, let us compute $RMSD\left(\hat{H}\right)$ using the formula of Eq. (11) using 60 points $\left(n=60\right)$ uniformly located on the beam. The respective values are presented in Table 3. One can conclude that in Variant A the best models are provided by SMS estimation, in Variants B and C – AMS estimation. Generally, AMS estimation provides the models for which the fit accuracy is almost the same in all variants.

Table 3. Accuracy of fit of the profiles in Variants A, B, and C

Method	$RMSD\left(\hat{H}\right)$ [mm]		
	Variant A	Variant B	Variant C
LS estimation	22.4	35.6	1.5
SMS estimation	0.2	2.3	4.1
AMS estimation	0.3	0.4	0.3

Considering the occurrence of outliers within the data, let the simulated sets be processed by applying robust M-estimation methods in the way applied in the previous example. The accuracy of the fit for both methods and all variants is presented in Table 4. All models determined are better fitted than the models of LS estimation, respectively. There is no doubt that the Huber method application provides the best models for M-estimation. It can deal with outliers of high magnitude much better than the Tukey method. The accuracy of the model determined by the application of the Huber method is like the accuracy of AMS estimation models in Variants A and B and lower in Variant C. In the paper context, AMS estimation can be an alternative to conventional robust methods when outliers have a high magnitude. It can provide better results when the outlier magnitude is low.

Table 4. Accuracy of fit of the profiles determined by M-estimation in Variants A, B, and C

Method	$RMSD\left(\hat{H}\right)$ [mm]		
	Variant A	Variant B	Variant C
Huber	0.2	0.3	1.1
Tukey	11.1	26.7	1.3

4. CONCLUSIONS

The paper concerns processing heterogeneous data from modern measurement techniques like LiDAR systems, including ALS and TLS. It might also concern other techniques providing big data observation sets, e.g., global navigation satellite systems (GNSS) observations. In the case of LiDAR data, heterogeneity of observations results from measuring different objects, namely the study object and "obstacles." Heterogeneity of the observation data might also result from different accuracy of observation groups or occurrence of outlying observations. Since modern techniques have become more popular nowadays, the problem of processing heterogeneous data is essential.

One of the possible ways to process heterogeneous data is the application of M_{split} estimation. The paper presents two numerical examples in that context. The first one concerns estimating two versions of the functional model parameters, which is a natural approach in M_{split} estimation. The observation sets contain two main groups of observations: terrain measurements and vegetation cover measurements. The obtained outcomes show that M_{split} estimation, especially AMS estimation, can provide better results than the conventional methods. AMS estimation shows its low sensitivity to outlying observations, and it overperforms the conventional robust methods. The second example concerns data, including the measurement of the study object and additional outlying observations. Hence, one is interested in estimating only one parameter version (describing the study object). The second solution of M_{split} estimation should be ignored as describing the location of outliers. This time AMS estimation also provides the best results that are similar in accuracy to conventional robust methods if the outliers are high magnitude. If data are disturbed by outliers of lower magnitude, then AMS estimation overperforms the traditional methods.

An alternative for M_{split} estimation would be data cleaning methods. One can say that in the second example in Variants A and B, it is easy to separate the measurements of the beam from outlying observations (measurements of other construction elements). As mentioned in the Introduction, such methods do not always succeed, e.g., it would be hard to separate regular observations from outliers in Variant C in the second example or Variants B and C in the first example. In that context, the application of M_{split} estimation seems advisable.

Generally, M_{split} estimation can be recommended to process heterogeneous data consisting of two (or more) observation groups for which the functional model parameters are estimated. The method in question can also be a good alternative to conventional robust methods, especially if outliers have low magnitude or high share within the observation set.

REFERENCES

Baselga S (2011) Nonexistence of rigorous tests for multiple outlier detection in least-squares adjustment. Journal of Surveying Engineering 137:109–112.

Baselga S (2007) Global optimization solution of robust estimation. Journal of Surveying Engineering 133:123–128.

Baselga S, Klein I, Suraci SS, Oliveira LC, Matsuoka MT, Rofatto VF (2021) Global Optimization of Redescending Robust Estimators. Mathematical Problems in Engineering 2021:1–13.

Błaszczak-Bąk W, Janowski A, Kamiński W, Rapiński J (2015) Application of the M_{split} method for filtering airborne laser scanning data-sets to estimate digital terrain models. International Journal of Remote Sensing 36:2421–2437.

Cabaleiro M, Riveiro B, Arias P, Caamaño JC (2015) Algorithm for beam deformation modeling from LiDAR data. Measurement 76:20–31.

Carrilho AC, Galo M (2019) Automatic object extraction from high resolution aerial imagery with simple linear iterative clustering and convolutional neural networks. ISPRS – International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences, XLII-2/W16:61–66.

Chen Z, Gao B, Devereux B (2017) State-of-the-art: DTM generation using airborne LIDAR data. Sensors 17:150.

Crespo-Peremarch P, Tompalski P, Coops NC, Ruiz LÁ (2018) Characterizing understory vegetation in Mediterranean forests using full-waveform airborne laser scanning data. Remote Sensing of Environment 217:400–413.

Czaplewski K, Wąż M, Zienkiewicz MH (2019) A novel approach of using selected unconventional geodesic methods of estimation on VTS areas. Marine Geodesy 42:447–468.

Ge Y, Yuan Y, Jia N (2013) More efficient methods among commonly used robust estimation methods for GPS coordinate transformation. Survey Review 45:229–234.

Gordon SJ, Lichti DD (2007) Modeling terrestrial laser scanner data for precise structural deformation measurement. Journal of Surveying Engineering 133:72–80.

Guo Y, Li Z, He H, et al (2020) A squared M_{split} similarity transformation method for stable points selection of deformation monitoring network. Acta Geodaetica et Cartographica Sinica 49:1419–1429.

Holst C, Burghof M, Kuhlmann H (2014) Modeling the beam deflection of a gantry crane under load. Journal of Surveying Engineering 140:52–59.

Janicka J, Rapiński J (2013) M_{split} transformation of coordinates. Survey Review 45:269–274.

Janicka J, Rapiński J, Błaszczak-Bąk W, Suchocki C (2020) Application of the M_{split} estimation method in the detection and dimensioning of the displacement of adjacent planes. Remote Sensing 12:3203.

Janowski A (2018) The circle object detection with the use of M_{split} estimation. E3S Web Conf 26:00014.

Janowski A, Rapiński J (2013) M-split estimation in laser scanning data modeling. Journal of the Indian Society of Remote Sensing 41:15–19.

Li J, Wang A, Xinyuan W (2013) M_{split} estimate the relationship between LS and its application in gross error detection. Mine Surveying 2:57–59.
Nowel K (2019) Squared $M_{split(q)}$ S-transformation of control network deformations. Journal of Geodesy 93:1025–1044.
Wiśniewski Z (2009) Estimation of parameters in a split functional model of geodetic observations (M_{split} estimation). Journal of Geodesy 83:105–120.
Wiśniewski Z (2010) $M_{split(q)}$ estimation: estimation of parameters in a multi split functional model of geodetic observations. Journal of Geodesy 84:355–372.
Wiśniewski Z, Zienkiewicz MH (2016) Shift-M_{split}* estimation in deformation analyses. Journal of Surveying Engineering 142:04016015.
Wyszkowska P, Duchnowski R (2019) M_{split} estimation based on L_1 norm condition. Journal of Surveying Engineering 145:04019006.
Wyszkowska P, Duchnowski R (2020) Iterative process of $M_{split(q)}$ estimation. Journal of Surveying Engineering 146:06020002.
Wyszkowska P, Duchnowski R, Dumalski A (2021) Determination of terrain profile from TLS data by applying Msplit estimation. Remote Sensing 13:31.
Zienkiewicz MH, Czaplewski K (2017) Application of square M_{split} estimation in determination of vessel position in coastal shipping. Polish Maritime Research 2(94):3–12
Zienkiewicz MH, Hejbudzka K, Dumalski A (2017) Multi split functional model of geodetic observations in deformation analyses of the Olsztyn castle. Acta Geodynamica et Geomaterialia 14:195–204.

BIOGRAPHICAL NOTES

Patrycja Wyszkowska – PhD, Assistant professor at Department of Geodesy, Institute of Geodesy and Civil Engineering, Faculty of Geoengineering, University of Warmia and Mazury in Olsztyn, Poland. Field of interest: estimation theory, robust estimation in geodetic computations, deformation analysis.
Robert Duchnowski – PhD habil., Head of Department of Geodesy, Institute of Geodesy and Civil Engineering, Faculty of Geoengineering, University of Warmia and Mazury in Olsztyn, Poland. Field of interest: estimation theory, robust estimation in geodetic computations, deformation analysis.

CONTACTS

PhD Patrycja Wyszkowska
PhD habil. Robert Duchnowski
Department of Geodesy, Institute of Geodesy and Civil Engineering, Faculty of Geoengineering, University of Warmia and Mazury in Olsztyn
Oczapowskiego 1 Street
Olsztyn
POLAND
Email: patrycja.wyszkowska@uwm.edu.pl, robert.duchnowski@uwm.edu.pl