Handling Outliers in Panel Data Models: A Robust Approach

ANABELA ROCHA[1,2], M. CRISTINA MIRANDA[1,2,3], MANUELA SOUTO DE MIRANDA[2]
[1]ISCA
University of Aveiro
Campus Santiago, 3810-193 Aveiro
PORTUGAL
[2]CIDMA
University of Aveiro
PORTUGAL
[3]CEAUL
University of Lisbon
PORTUGAL

Abstract: Real-world data often violate the conditions assumed by classical estimation methods. One reason for this failure may be the presence of observations with a low probability of belonging to the same distribution as the majority of the data, known as outliers. Outliers can appear in different forms, such as casewise and cellwise outliers. The results of classical estimation methods, particularly those based on least squares, can be seriously affected by the presence of any type of outlier. Panel data modeling is applied in various fields, including economics, finance, marketing, biology, environmental studies, healthcare, and more. The estimation of these models is typically performed using classical methods. In this paper, we consider the random effects panel data model and propose a robust method to estimate the parameters of this model. To evaluate the performance of the proposed robust estimation method compared to the classical estimation method, we conducted a Monte Carlo simulation study. Additionally, we illustrate the proposed methodology by applying it to estimate a model based on a real panel data set.

Key-Words: - Cellwise outliers, Casewise outliers, Panel data model, Robust methods, Simulation.

Received: April 4, 2024. Revised: September 6, 2024. Accepted: October 9, 2024. Published: November 25, 2024.

1 Introduction

Statistical models always constitute simplifications of real-life problems. Consequently, in real data sets, it is common to encounter elements that deviate from the pattern followed by the majority of the data. These elements, known as outliers, can arise for various reasons and may present challenges for statistical analysis. In fact, outliers can seriously distort a statistical analysis, but they can also provide important information and are therefore warrant closer examination. It is important to have methods that allow us to detect outliers that may exist within data sets and to use robust techniques that minimize the impact of these outliers on the overall results.

Identifying outliers is crucial in data analysis as they can significantly affect conclusions and interpretations. Outliers are data points markedly different from others in a data set, because they are highly above (or down bellow) the rest of the data set. They can arise due to measurement errors, unique events, or anomalies. In panel data—where observations are collected over time from multiple entities such as individuals, firms, or regions—detecting outliers becomes more complex. This complexity stems from the need to differentiate between genuine anomalies and meaningful variations across entities or time periods. Failure to properly identify and handle outliers in panel data can lead to biased estimates, misleading trends, and inaccurate predictive models. Therefore, robust methods tailored to panel data, such as robust statistical measures or techniques accounting for temporal and cross-sectional dependencies, are essential to ensure reliable analyses and valid conclusions.

Several robust estimation methods for Panel Data Model (PDM) have been proposed. However, these methods are not robust against all types of outliers. In the last years, two distinct types of outliers have gained attention: casewise outliers and cellwise outliers. While robust methods against cellwise outliers have been developed and published, they have not yet been specifically tailored for PDM estimation. The method proposed in this work provides a robust approach for estimating PDM, addressing both casewise and cellwise outliers.

In this paper the authors consider the random effects estimator as defined in the usual Econometrics literature [1], [2], [3] and replace the covariance matrix by a more robust version. At the same time, they consider a recent scheme for identifying outliers and to control the effects of those

outliers in the rest of the data [4]. A Monte Carlo simulation study is performed to assess the validity of the implemented procedure. In face of the good results measured in terms of the Root Mean Squared Error (RMSE), a set of real data is also used to compare the results between the classical FGLS estimator and the new proposal, present in this paper.

This paper is organized as follows: Section 1 provides a brief introduction. Section 2 discusses some of the robust methods considered. In Section 3, the focus shifts to panel data, introducing the classic Feasible Generalized Least Squares (FGLS) estimator, as well as the Robust Feasible Generalized Least Squares (RFGLS) estimator proposed in this work. Section 4 presents a simulation study showing the strong performance of the methodologies used. In Section 5, these procedures are tested on a real data set, specifically the Grunfeld data [5]. Finally, Section 6 concludes the paper by summarizing the main findings.

2 Robust Methods

Let us consider a data set in a matrix form, in which the rows are the cases, and the columns are the variables. Different types of outliers may occur. Traditionally, outlier refers to a case or a row of the data matrix. This is called a casewise outlier. [6] proposed the identification of a cellwise outlier. These occur when most of the data cells in a row are similar, but some of them are atypical.

[4] give an illustration of this phenomena which we include in Figure 1, and show that a small percentage of cellwise outliers can lead to many casewise outliers.

Classical estimation methods, in particular those based on the least squares method, may be seriously affected with the presence of outliers. Detecting outliers in datasets is critical, but visual inspection becomes challenging in the context of multivariate data. Robust fitting methods which are less sensitive to casewise outliers and allow to detect those outliers can be seen in [7]. Recent work, such as [8], has focused on identifying cellwise outliers and addressing them in the estimation and fitting processes. Among the available proposals, we refer to the Univariate-and-bivariate filter (UBF) [9], [10]. The univariate filter flags cellwise outliers by comparing the standardized empirical distribution of each marginal with a high quantile of the standard normal distribution. The bivariate filter flags casewise outliers by comparing the squares of the pairwise robust Mahalanobis distances with a high quantile of a chi-square with two d.f. distribution. A cell is additionally flagged when the number of the flagged pairs exceeds a large quantile of the binomial model, considering that the number of the flagged pairs associated with each cell approximately follows a binomial model.

We also refer to the the cellMCD method, proposed by [11]. This method is a cellwise robust version of the minimum covariance determinant (MCD) estimator of Rousseeuw [12], which is a covariance matrix estimator that is robust against casewise outliers.

This robust method for location and scale estimation identifies the subset of h sample observations that minimizes the determinant of the sample covariance matrix, where h is an integer greater than or equal to half the sample size n. The location MCD estimator corresponds to the center (mean) of that subset, and the scale MCD estimator corresponds to matrix that defines its shape (covariance matrix). The MDC estimates are not much affected when samples contain fewer than n-h outlying cases, but is not robust against cellwise outliers. The cellMCD method overcome this flaw, resulting of the minimization of an objective function with good breakdown properties and consistency.

Finally, among the robust regression methods, we refer to the Least Trimmed Squares (LTS) estimator [12]. This robust regression method estimates model parameters by minimizing the sum of the h smallest squared residuals. By doing so, it excludes the n-h largest absolute residuals from the estimation process, effectively tolerating n-h outliers. We highlight these robust methods as they will be used later in this work, when formulating the proposed robust estimation method.

3 Panel Data

A panel data set consists in a number of observations from a set of variables for different units (e.g. countries, firms, regions, individuals) which are

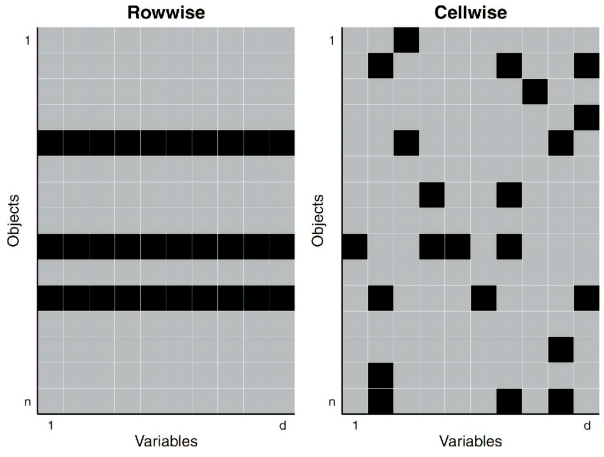

Fig. 1: Rousseeuw and Bossche illustration [4]

collected over several periods of time (e.g. daily, weekly, monthly). The number of time periods is usually small in comparison with the number of units being studied. This type of data can reveal more information than observing the same variables across different units at a single point in time (cross-sectional data) or observing the same variables in a single unit over multiple time periods (time series data). The statistical analysis of panel data allows to identify and to measure effects that would not be identified in an cross-sectional analysis nor in a time-series analysis. Let $\mathbf{X}$ and y be observable random variables, and let μ be an unobservable random variable. The main interest is to study the partial effects of the observable explanatory variables x_j in the dependent variable y, which can be represented by the regression equation (1).

$$y_{it} = \mathbf{X}_{it}\boldsymbol{\beta} + u_{it}, i = 1, ..., N; t = 1, ..., T, \quad (1)$$

with $u_{it} = (\mu_i + \nu_{it})$; the index i refers to units (like countries, firms, regions, individuals) and the index t refers to time periods (like day, week, month, year); $\boldsymbol{\beta}$ is a vector of K parameters; $\mathbf{y}_{it}$ is the observation i of the dependent variable y in the period t; $\mathbf{X}_{it}$ is the observation i of the K explanatory variables in the period t; μ_i is the unobservable individual effect and ν_{it} represent the remainder disturbance and are usually called idiosyncratic errors.

According to the assumptions assumed for the model defined in (1), we may have a fixed effects model or a random effects model. To consider the existence of correlation between errors over time, or between firms (or individuals), we should consider the Fixed Effects (FE) or the Random Effects (RE) model. The difference upon these two models is the existence (FE) or non existence (RE) of correlations between the unobservable individual effect μ_i and the explanatory variables.

3.1 Random Effects Model

In this work, we study the Random Effects (RE) model. To obtain unbiased and consistent estimators, the model requires the following assumptions [1]:

- μ_i are independent and identically distributed (IID) random variables with zero mean and variance σ_μ^2;
- ν_{it} are IID random variables with zero mean and variance σ_ν^2;
- μ_i and ν_{it} are independents;
- X_{it} are independent of the μ_i and $\nu_{it}, \forall i, t$.

For individual i, the panel data model defined in Equation (1) can be represented by the Equation (2)

$$\mathbf{y}_i = \mathbf{X}_i\boldsymbol{\beta} + \mathbf{u}_i. \quad (2)$$

In vector form, with all observations stacked, the same equation can be rewritten using the expression (3).

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{u}, \quad (3)$$

and

$$\boldsymbol{\Omega} = E(\mathbf{u}\mathbf{u}') = \mathbf{I}_N \otimes (\sigma_\mu^2 \mathbf{J}_T + \sigma_\nu^2 \mathbf{I}_T) = \mathbf{I}_N \otimes \mathbf{V}, \quad (4)$$

with $\mathbf{J}_T$ being a matrix of ones of order T. In this case, we assume homoscedasticity and serial correlation over time only between the errors of the same individual, i.e., the $\mathbf{V}$ matrix has a structure according to (5).

$$\mathbf{V} = \begin{bmatrix} \sigma_\mu^2 + \sigma_\nu^2 & \sigma_\mu^2 & \dots & \sigma_\mu^2 \\ \sigma_\mu^2 & \sigma_\mu^2 + \sigma_\nu^2 & \dots & \sigma_\mu^2 \\ \dots & \dots & \dots & \dots \\ \sigma_\mu^2 & \sigma_\mu^2 & \dots & \sigma_\mu^2 + \sigma_\nu^2 \end{bmatrix}_{(TxT)} \quad (5)$$

3.2 Feasible Generalized Least Squares Estimator

For the RE model, the Feasible Generalized Least Squares (FGLS) obtained estimator may be expressed by (6).

$$\hat{\boldsymbol{\beta}}_{FGLS} = \left(\mathbf{X}'\hat{\boldsymbol{\Omega}}^{-1}\mathbf{X}\right)^{-1}\mathbf{X}'\hat{\boldsymbol{\Omega}}^{-1}\mathbf{y}, \quad (6)$$

where $\hat{\boldsymbol{\Omega}}$ represents an estimate of the covariance matrix of the errors of the model. Estimating the $\boldsymbol{\Omega}$ matrix is equivalent to estimating the $\mathbf{V}$ matrix and this requires the estimation of the variance components σ_μ^2 and σ_ν^2.

The best quadratic unbiased estimators of the variance components in Equation (4) are given by the following expressions:

$$\hat{\sigma}_\nu^2 = \frac{\sum_{i=1}^{N}\sum_{t=1}^{T}(u_{it} - \bar{u}_{i.})^2}{N(T-1)}; \quad (7)$$

$$\hat{\sigma}_\mu^2 = \frac{(\hat{\sigma}_1^2 - \hat{\sigma}_\nu^2)}{T}; \quad (8)$$

$$\hat{\sigma}_1^2 = T\frac{\sum_{i=1}^{N}\bar{u}_{i.}^2}{N}, \quad (9)$$

where $\hat{\boldsymbol{\Omega}}$ represents an estimate of the errors of the model covariance matrix. In this work, we replace the sample means and variances that appear in these expressions by robust location and scale estimates, respectively.

3.3 Robust FGLS Estimator

The problems mentioned above also commonly occur in real panel data, namely, the failure to fulfill the assumed assumptions, or the presence of atypical observations. Classical estimation methods of PDM, in particular FGLS, may be seriously affected with the presence of outliers. It's important to detect the existence of outliers in panel data, and the use of robust estimation methods in that case can make a difference in terms of the results accuracy, as they are less affected by the outliers behaviour.

Some robust procedures have been proposed for PDM, like [13], [14], [15], [16], where the authors adapted robust regression methods [7] to PDM.

The methods early proposed for PDM are robust against casewise outliers, but not against cellwise outliers. The new robust estimation method proposed is robust against both types of outliers, and constitutes a good alternative to estimate PDM when data include casewise or cellwise outliers.

In this paper, the authors propose a RFGLS (Robust Feasible Generalized Least Squares) estimator, resulting from the implementation of robust procedures in the three steps that support the process of obtaining the FGLS estimator. This proposal results from a refinement of the method proposed by [17], and it is a consequence of the application of more robust methods in the RFGLS computation process. In this case, the authors have selected the median per variable when replacing the identified outliers in the UBF, and have used the cellMCD method, which is robust to cellwise and casewise outliers.

The RFGLS algorithm can be summarized in the following subsection:

3.3.1 Robust Feasible Generalized Least Squares Algorithm

1. Estimate the Pooled Model parameters using Least Trimmed Squares (LTS) estimator and compute the residuals.
2. Estimate the errors covariance matrix using the robust covariance matrix estimator cellMCD applied to the residuals obtained in the previous step.
3. Filter the original data matrix using the univariate-and-bivariate filter (UBF).
4. Obtain the cleaned data matrix, replacing each identified outlier by UBF in the former step by the median of the corresponding variable.
5. Estimate the model parameters by FGLS from the cleaned data matrix obtained at the fourth step and using the robust estimated covariances matrix obtained at the second step.

4 Simulation Study

For the evaluation of the performance of the proposed robust estimator, RFGLS, the authors run a simulation study. A data set was randomly generated, and next suffered a contamination process; a number of outliers was included in the simulated samples; this was done in two distinct ways and considering different percentages of contamination. In the simulation settings the authors followed the papers [13], [14] and [18].

All the calculations were carried out with the R project [19]. The authors used the packages *plm*, specific for analysing panel data, and the *robustbase*, *GSE* and *cellWise* for some of the robust methods implemented and cellwise outlier detection.

4.1 Settings

The explanatory variables values were generated from a multivariate (dimension three) standard normal. For the parameter vector, the values were generated with $\boldsymbol{\beta} = (-1, 0, 1)$ and μ_i, according with a $N(0, 1)$ distribution. The errors values were generated according with a $N(0, 1)$ distribution. The dependent variables values were obtained according to the RE model, defined *e.g.* in [1].

The panel data sets were generated with 240 observations, resulting from two scenarios for the dimensions: ($N = 8$ and $T = 30$), and ($N = 12$ and $T = 20$).

4.2 Contamination

In the sample generation process, three different values of percentages of contamination were considered to insert into the samples: 0%, 5%, and 10%. The contamination process was completely random over all observations of the panel data, and it was introduced including outliers in two different ways as follows:

1. the contamination is led only on y (to originate vertical outliers), by adding to some of the y initially generated, a term generated according with $N(50, 1)$;
2. the contamination is led on y and x (to originate bad leverage points), by replacing the explanatory variables values, corresponding to the observations already contaminated in y, by points coming from a k-variate $N(10, I)$ distribution, with $k = 3$.

A total of $M = 100$ replications for each of the 10 sampling schemes was conducted, resulting in a total of 20 scenarios and 200 runs.

Table 1. RMSE for FGLS and RFGLS, $T = 30$, $N = 8$ and contamination 0%, 5% and 10%

Contamination	FGLS	RFGLS
0%	0.15	0.25
5% in y	1.26	0.23
5% in y and x	2.72	0.22
10% in y	1.73	0.26
10% in y and x	2.82	0.21

Table 2. RMSE for FGLS and RFGLS, $T = 20$, $N = 12$ and contamination 0%, 5% and 10%

Contamination	FGLS	RFGLS
0%	0.16	0.25
5% in y	1.11	0.23
5% in y and x	2.71	0.21
10% in y	1.67	0.25
10% in y and x	2.81	0.19

4.3 Estimation and Performance Assessment

In each run, we estimate the $\boldsymbol{\beta}$ coefficient of the model using both methods, *i.e.*, the FGLS and the RFGLS estimators. The performance of the two approaches was evaluated based on the Root Mean Squared Error (RMSE) criteria, which was computed RMSE according to the expression (10):

$$RMSE = \sqrt{\frac{1}{M}\sum_{j=1}^{M}\left|\left|\hat{\boldsymbol{\beta}}^{j} - \boldsymbol{\beta}\right|\right|^{2}}, \qquad (10)$$

for $\boldsymbol{\beta}$ referred in the simulation settings (4.1) and the $\boldsymbol{\beta}$ estimates were obtained for each of the $M = 100$ replicates, with each estimator, corresponding to a different sampling scheme. The estimator with the best performance corresponds to the estimator that presents the lowest RMSE value, as this value provides a measure of the estimation error.

4.4 Results

Table 1 contains the RMSE values of the FGLS and the RFGLS estimators in case $T = 30$, $N = 8$ and contamination levels 0%, 5% and 10%.

The RMSE values are always smaller for the RFGLS estimator, except in the case where there is no data contamination. This means that the robust estimator generates more accurate estimates in all the contamination situations considered. We also notice that in presence of bad leverage points, corresponding to the case of contamination on y and x, the results obtained by FGLS are especially affected in a negative way, while the RFGLS continues to present particularly good results. The only scenario for which the RFGLS has not a higher performance in the case with cleaned samples, without outliers.

In Table 2, RMSE values of FGLS and RFGLS estimators are presented for $T = 20$, $N = 12$ and contamination levels 0%, 5% and 10%:

The results given in Table 2 were obtained for data panels with $T = 20$ and $N = 12$, and are very similar to those in Table 2, for panels with $T = 8$ and $N = 30$. It is also possible to observe in this case, that the produced estimates with RFGLS are more precise, with lower RMSE than those obtained with FGLS, for all contamination cases considered.

We can summarize and conclude that for all the considered scenarios of contamination, the RMSE values for the robust estimator are always lower than the ones obtained for the classical estimator. So the robust estimator results improve, as expected, in the presence of outliers. Without contamination, the efficiency of the robust estimator is not as good as the one of the classic estimator.

5 Grunfeld Data

The Grunfeld data is a well known panel data set among econometrics researchers. It is a panel data containing annual observations for US companies, with a total of 220 observations (11 companies $\times$20 years), and corresponds to the values of the following variables (in dollars with reference to the year 1947):

- *invest*, corresponding to Gross investment;
- *value*, corresponding to Market value;
- *capital*, corresponding to Stock of plant and equipment;
- *firm*, taking the values General Motors, US Steel, General Electric, Chrysler, Atlantic Refining, IBM, Union Oil, Westinghouse, Goodyear, Diamond Match, American Steel;
- *year*, taking the values from 1935 to 1954.

To describe the dependence relationship of the investment in relation to the value and capital, Grunfeld formulated the model equation (11) [20]:

$$invest_{it} = \beta_0 + \beta_1 value_{it} + \beta_2 capital_{it} + u_{it}. \qquad (11)$$

An exploratory graphical analysis shows the existence of atypical observations of various types. We can see casewise outliers, corresponding to companies whose values are very different from those of the rest of the firms; and also cellwise outliers, related to years in which the values of some companies are very different from those of the

remaining companies. Therefore, it seems that is more appropriate to estimate the model parameters using a robust estimator than using the classical estimators. Figure 2 shows the atypical behaviour of two of the firms, with values high above the rest of the companies.

Figure 3 shows how misleading would be a fit for the Value variable in this case, without accounting for this differentiation.

To illustrate the application of both considered estimators, FGLS and RFGLS, the authors estimate the model parameters using both methods. To compare the performance between the two methods, the prediction performance was used, defined by the root mean squared error of prediction (RMSEP) calculated over the set of uncontaminated cases, defined in (12). This evaluation criterion was proposed in [21] to compare the relative performance of different approaches.

Since the RMSEP corresponds to a mean prediction error, the method with the best performance is the one with the lowest RMSEP value, since on average, it presents predictions closest to the observed values. The clean residuals were also produced, obtained for uncontaminated cases, and a descriptive analysis of these waste values was carried out. The lower the residual

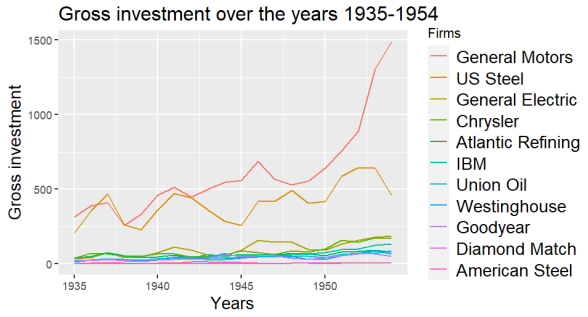

Fig. 2: Grunfeld data: two firms are very different from the rest

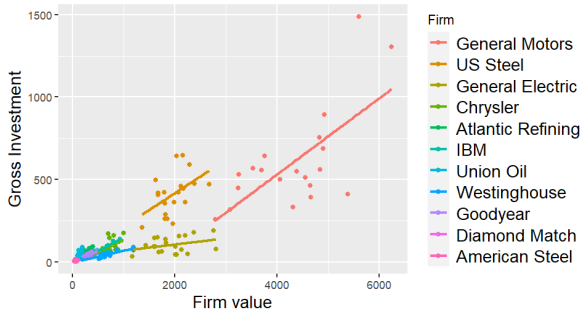

Fig. 3: Grunfeld data shows the need of fit readjustments

Table 3. Estimates and RMSEP for FGLS and RFGLS with Grunfeld data.

	FGLS	RFGLS
value	0.11	0.07
capital	0.31	0.11
RMSEP	63.01	18.89

Table 4. Descriptive statistics obtained for clean residuals.

	FGLS	RFGLS
mean	48.09	13.16
median	34.58	7.73
SD	40.85	14.96
MAD	37.69	7.76

values, the better the method will perform.

$$RMSEP = \sqrt{\frac{1}{Nc}\sum_{i\in I}(\hat{y}_i - y_i)^2}, \qquad (12)$$

where I contains the indices of clean, uncontaminated cases and Nc is the number of uncontaminated cases.

Table 3 contains the parameter estimates obtained with both methods and the RMSEP values computed for both methods.

The root mean squared error of prediction (RMSEP) values in Table 3 shows that the robust method performs better. The RMSEP value obtained with the robust method RFGLS is smaller than the obtained with the classical method FGLS.

Table 4 contains the values of some descriptive statistics obtained for the set of clean residuals, that is, calculated from the uncontaminated observations or cases. To evaluate the residual location, the mean and the median were computed, and to evaluate the residual scatter, the standard deviation (SD) and the median absolute deviation (MAD) were determined.

Table 4 shows that the values of the clean residuals, obtained for uncontaminated cases, present lower location (mean and median values) and lower dispersion (SD and MAD values) for RFGLS method than for the FGLS method. Therefore, we may conclude that the robust method leads to more accurate results than the classical method, as it leads to smaller residuals, and with less variability.

The illustration with the Grunfeld data, which contains outliers, shows that the RFGLS robust method proposed in this paper performs better than its classical version, FGLS. This conclusion is supported by the fact that the robust method presents a lower error value and is associated with smaller residuals. This indicates that the robust method allows us to obtain more accurate predictions, taking into account these performance evaluation criteria.

6 Final Comments

Panel data is a suitable representation for the most diverse areas of knowledge. Real panel data often contain outliers and violate the assumptions usually assumed in the Subsection (3.1), e.g., non normal distributed, different variances or means. Robust methods are therefore recommended for this type of data analysis. The authors propose a robust estimator for panel data model in the present paper, which results from a robustification of the FGLS estimator, called RFGLS. This process involved the application of robust methods to cellwise and casewise outliers at various stages of the FGLS calculation process. In particular, the LTS robust regression method, the cellMCD covariance matrix robust estimation method, and the univariate and bivariate UBV filter were applied to detect and flag outliers. The RFGLS estimator performed well with contaminated simulated data. The authors carried out a simulation study to compare the performance of the FGLS estimator with the RFGLS estimator, considering several simulation scenarios. The proposed robust estimator RFGLS improve regarding to FGLS as expected in the presence of several type of outliers. RFGLS performs better than FGLS because RFGLS present lower RMSE than FGLS in the presence of vertical outliers and leverage points, and for all dimensions and percentages of contamination considered. RFGLS produces particularly good estimates for panels of data with bad leverage points. The RFGLS estimator did not perform as well without contaminated data, as expected.//

The authors have also illustrated the application of the RFGLS estimator when estimating parameters for a model fitted to a real data set, initially proposed by Grunfeld to model investment. Also for the Grunfeld data, which is a panel data that contains outliers, the robust estimator performs better than the classical estimator. The robust estimated model is less affected by the identified outliers then the classical estimated model. The authors plan to continue the present work performing more simulations with different scenarios. This will allow a better evaluation of the robustness properties of the proposed estimator. Different contamination in order to generate samples with casewise outliers, cellwise outliers and outliers concentrated in certain groups (concentrated contamination) are other cases that would also be interesting to analyse. Furthermore, we suspect that including a robust regression method in the last step of the RFGLS algorithm may further improve the robustness properties of the robust estimator.

Acknowledgment:
Research partially supported through the Portuguese Foundation for Science and Technology (FCT - Fundação para a Ciência e a Tecnologia), by the Center for Research and Development in Mathematics and Applications (CIDMA),UIDB/04106/2020, doi.org/10.54499/UIDB/04106/2020, doi.org/10.54499/UIDP/04106/2020,and Centre of Statistics and its Applications (CEAUL), within project UIDB/00006/2020 (https://doi.org/10.54499/UIDB/00006/2020).

Declaration of Generative AI and AI-assisted technologies in the writing process:
During the preparation of this work the authors used *Grammarly* for language editing. After using this service, the authors reviewed and edited the content as needed and take full responsibility for the content of the publication.

References:

[1] J. M. Wooldridge. *Econometric Analysis of Cross Section and Panel Data*. The MIT Press, 2nd edition, 2010.

[2] B. H. Baltagi and J. M. Griffin. Gasoline demand in the OECD: An application of pooling and testing procedures. *European Economic Review*, 22(2):117–137, 1983.

[3] W. H. Greene. *Econometric Analysis Global Edition* . Pearson, 8th edition, 2020.

[4] P. J. Rousseeuw and W. Van Den Bossche. Detecting Deviating Data Cells. *Technometrics*, 60(2):135–145, apr 2018.

[5] C. Kleiber and A. Zeileis. The Grunfeld Data at 50. *german Economic review*, 11:404–417, 2010.

[6] A. Fatemah, Van A. Stefan, Victor J. Y., and Ruben H. Z. Propagation of outliers in multivariate data. *The Annals of Statistics*, 37(1):311–331, 2009.

[7] R. A. Maronna, R. D. Martin, and V. J. Yohai. *Robust Statistics: Theory and Methods*. Wiley, jun 2006.

[8] J. Raymaekers and P. J. Rousseeuw. Challenges of cellwise outliers. *Econometrics and Statistics*, 2024.

[9] C. Agostinelli, A. Leung, V. J. Yohai, and R. H. Zamar. Rejoinder on: Robust estimation of multivariate location and scatter in the presence of cellwise and casewise contamination. *Test*, 24(3):484–488, sep 2015.

[10] A. Leung, V. Yohai, and R. Zamar. Multivariate location and scatter matrix estimation under cellwise and casewise contamination. *Computational Statistics and Data Analysis*, 111:59–76, jul 2017.

[11] J. Raymaekers and P. J. Rousseeuw. The cellwise minimum covariance determinant estimator. *Journal of the American Statistical Association*, 0(0):1–12, 2023.

[12] P. J. Rousseeuw. Least Median of Squares Regression. *Journal of the American Statistical Association*, 79(388):871, dec 1984.

[13] M. C. Bramati and C. Croux. Robust estimators for the fixed effects panel data model. *Econometrics Journal*, 10(3):521–540, 2007.

[14] M. Aquaro and P. Čížek. One-step robust estimation of fixed-effects panel data models. *Computational Statistics and Data Analysis*, 57(1):536–548, 2013.

[15] G. Dhaene and Y. Zhu. Median-based estimation of dynamic panel models with fixed effects. *Computational Statistics and Data Analysis*, 113(C):398–423, sep 2017.

[16] A. Ji, B. Wei, and L. Xu. Robust estimation of panel data regression models and applications. *Communications in Statistics - Theory and Methods*, 52(21):7647–7659, November 2023.

[17] A. Rocha and M. C. Miranda. Robust Estimation for the Random Effects Panel Data Models. In *New Frontiers in Statistics and Data Science (SPE 2023, Guimarães, Portugal), in press*. Springer, 2024.

[18] M. Aquaro and P. Čížek. One-step robust estimation of fixed-effects panel data models. *Computational Statistics and Data Analysis*, 57(1):536–548, jan 2013.

[19] R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2022.

[20] D. N. Gujarati. *Basic Econometrics* . Tata McGraw Hill, 4th edition, 2004.

[21] P. Filzmoser, S. Höppner, I. Ortner, S. Serneels, and T. Verdonck. Cellwise robust m regression. *Computational Statistics and Data Analysis*, 147:106944, 2020.

Contribution of Individual Authors to the Creation of a Scientific Article (Ghostwriting Policy)

The three authors were responsible for the conceptualization of the paper and for the writing/editing of the manuscript. Anabela Rocha and Manuela Souto de Miranda proposed the algorithm. Anabela Rocha and M. Cristina Miranda performed the formal analysis and applied the methodology. M. Cristina Miranda produced the code for the analysis and visualization.

Sources of Funding for Research Presented in a Scientific Article or Scientific Article Itself

Research partially supported through the Portuguese Foundation for Science and Technology (FCT - Fundação para a Ciência e a Tecnologia)

Conflict of Interest

The authors have no conflicts of interest to declare.