Socio-Economic Planning Sciences 95 (2024) 101992

Contents lists available at ScienceDirect

Socio-Economic Planning Sciences

journal homepage: www.elsevier.com/locate/seps

The robustification of distance-based linear models: Some proposals

Eva Boj [a,*], Aurea Grané [b]

[a] Departament of Economic, Financial and Actuarial Mathematics, University of Barcelona, Avda. Diagonal 690, Barcelona, Spain
[b] Statistics Department, Universidad Carlos III de Madrid, C/ Madrid 126, 28903 Getafe, Spain

ARTICLE INFO

Keywords:
dblm
dbstats
Mixed-type data
Outliers
Robust distance
R

ABSTRACT

In this work tailor robust metrics are proposed to be used in the predictors' space of distance-based predictive models. The first proposal is a robust version of Gower's distance, which takes into account the correlation structure of the data. The second one is a rather complex metric, constructed via Related Metric Scaling, which is able to discard redundant information coming from different sources. Another novelty is the proposal of a distance-based trimming statistic to robustify the metrics. The performance of the models based on new robust metrics is evaluated through a simulation study and compared to those based on Euclidean, Gower's and generalized Gower's metrics in the presence of outliers in several datasets of multivariate heterogeneous data. Mean squared error (also median and standard deviation) are used to evaluate the effectiveness in the prediction of responses. Finally, two applications in the areas of sustainable transport and finance and banking are provided in order to illustrate the predictive power of these models. Computations are made using the `dbstats` package for R.

1. Introduction

Methods relying on distances or similarities between sample units have a rich and longstanding tradition in statistics. Among these, cluster analysis and multidimensional scaling (MDS) [1] are widely employed. MDS serves as a multivariate dimensionality reduction technique that proves effective when information regarding the data is presented in the form of an inter-individual distance matrix. Originating from the metric version of MDS, the distance-based linear model (DB-LM) was introduced by [2] and subsequently expanded upon in works such as [3–9], and [10]. DB-LM is a prediction tool which can be applied to qualitative or mixed explanatory variables while keeping compatibility with ordinary regression by weighted least squares (WLS), which appears as a particular case when the Euclidean distance is used among individuals. The model projects the vector of continuous responses onto a Euclidean space obtained by MDS from the observed predictors, which are nonlinearly mapped into a set of latent, i.e., non-observed, variables in this space.

The model allows any distance among predictors, whose choice depends on the specific problem and the nature of the data. In this work, we focus on data of mixed type, which commonly arise in contexts such as economics, health, finance, marketing or sociodemographic surveys, among others. In such situations, when the information comes from categorical and numerical variables, the classical distance measure to be considered is Gower's distance [11], and thus DB-LM was traditionally built from Gower's metric (which, form now on, will be considered the benchmark model). However, this measure presents limitations in handling outliers and heavy-tailed data, which can lead to biased results. To overcome this drawback, in [12] new robust metrics were proposed for MDS and clustering purposes. We refer to them as robust generalized Gower's (G-Gower) and robust Related Metric Scaling (RelMS). In this paper, a new distance-based trimming statistic is used to robustify these metrics, and several trimming thresholds are evaluated. These new proposals are able to deal with outliers and variable redundancy. That is, to some extent variable selection is implicit in the construction of the metric. So far, robust metrics have not been used and studied in distance-based prediction models. Thus, the main aim of this paper is to robustify the DB-LM using these robust proposals.

To achieve these goals R code is developed implementing the formulation of the new proposals to allow the calculation of an object of type `dist` (or `D2` and `Gram`) following the standards of the `dbstats` package [9] for R [13]. The performance of the new metrics is evaluated in the context of distance-based prediction and compared to those of classical Euclidean and Gower's, and G-Gower's by means of the mean square error (MSE). In particular we focus on the DB-LM which is fitted through the function `dblm` of the `dbstats` package for R.

To evaluate the effectiveness of the predictions, Monte Carlo experiments are performed with three simulated mixed-type datasets with

* Corresponding author.
E-mail addresses: evaboj@ub.edu (E. Boj), aurea.grane@uc3m.es (A. Grané).

https://doi.org/10.1016/j.seps.2024.101992
Received 29 January 2024; Received in revised form 2 May 2024; Accepted 11 June 2024
Available online 18 June 2024

different number of predictors and different correlation/association structure among them. Outlier contamination by predictor's type is introduced, giving rise to twelve scenarios, and four different types of response variable are considered (contaminated/ uncontaminated, linear/ non linear). Distance models based on different metrics, such as Euclidean, Gower's, G-Gower's, robust G-Gower's and robust RelMS, are fitted to each of the datasets and next their performance in the prediction of the responses is evaluated by cross-validation procedures. The final goal of the paper is to conclude whether the DB-LM based on robust metrics is competitive with respect to existing distance-based models in the presence of anomalous data.

Finally, two applications on real datasets are included to illustrate the performance of the new robust DB-LM. Both datasets include anomalous observations and correlation/association among predictors. The first application is in the area of sustainable transport, where the aim is to predict the bike sharing demand in Capital Bikeshare program, which operates in the District of Columbia, Arlington County, VA, and the City of Alexandria, VA, from a mixed-type set of predictors concerning renting details and weather conditions (obtained from Ronald Reagan Washington National Airport). This dataset was collected by [10] from different sources and DB-LM with Gower's metric was proven to be very effective in front of other competitors. The second application is in the area of finance and banking, and is devoted to motor insurance. Data come from a study conducted by a committee on risk premiums in automobile insurance in Sweden, and in this case, the objective is to predict claim severity regarding a set of weighted mixed-type predictors related to the kilometers traveled, specific car makes and bonus. This dataset was analyzed in [8] were DB-LM with Gower's metric outperformed the classical Euclidean one. In both applications the prediction power of DB-LM with robust proposals is compared to those of other metrics such as Euclidean, Gower's and G-Gower's by cross-validation procedures.

The results of the simulation study and the analysis of two real datasets lead us to conclude that the performance in the prediction of the responses of DB-LM with robust proposals outperforms those of classical Gower's (which has been used up to date when working with mixed-type data), Euclidean and G-Gower's in the presence of anomalous data. In previous studies, it was seen that DB-LM with Gower's metric outperformed other predictive models (see [5–8,10]). Thus, we believe that using DB-LM with a robust metric is also a good alternative to these models.

The paper proceeds as follows: Section 2 contains general context notation, the definition of DB-LM, and distance matrices for predictors among which new robust proposals can be found. Section 3 contains the simulation study with a description of the scenarios considered for models' evaluation under anomalous individuals and the models' comparison in the prediction of responses. The predictive power of these models is illustrated on two real datasets in Section 4 and conclusions are given in Section 5. Additional results are included in Appendices A–C.

2. Methodology

2.1. The distance-based linear model

The concept of DB-LM consists of using the principal coordinates resulting from the MDS applied to a matrix of inter-individual distances as explanatory variables within a linear regression model. Just as the standard linear model has been extended to the generalized linear model (GLM), local linear regression or nonparametric versions of the GLM, the DB-LM can be extended as well. In this regard, [7] introduced local DB-LM, a nonparametric prediction technique that extends the DB-LM. In addition, two further extensions were introduced in [8]: distance-based generalized linear models (DB-GLM) and its nonparametric version (local DB-GLM) through local likelihood. These models, among others, are implemented in the `dbstats` package [9] for R. In particular the DB-LM can be fitted by means of the `dblm` function (we refer to the package help for the detail).

2.1.1. Preliminaries

Let $\Omega = \{\Omega_1, \dots, \Omega_n\}$ be a random sample of individuals from a given population, for which the value of a random quantitative response variable $Y \in \mathbb{R}$ is observed, that is, $\mathbf{Y} = (Y_1, \dots, Y_n)^\top$. For each $i = 1, \dots, n$, consider $w_i \in (0,1)$ a constant positive weight for Ω_i and let $\mathbf{w} = (w_1, \dots, w_n)^\top$ be the $n \times 1$ weight vector, such that $\mathbf{1}^\top \cdot \mathbf{w} = 1$, where $\mathbf{1}$ represents the $n \times 1$ vector of ones.

We consider a distance function $\delta(\cdot, \cdot)$ defined on the set Ω. Let $\mathbf{\Delta} = \left(\delta^2(\Omega_i, \Omega_j)\right)_{1 \le i,j \le n}$ be the $n \times n$ matrix of pairwise squared distances between individuals of Ω. A specific scenario arises when individuals in Ω are characterized by a set of variables which may encompass a combination of quantitative and qualitative measurements or unconventional quantities like character strings or functions. The distance δ can be expressed as a function of the variables. In this study we consider a matrix $\mathbf{Z}$ of p mixed predictor variables and we denote by $(\mathbf{z}_1, \dots, \mathbf{z}_n)^\top$ their corresponding measurements, where $\mathbf{z}_i \in \mathbb{R}^p$ for $i = 1, \dots, n$. Then the squared distance matrix $\mathbf{\Delta} = (\delta^2(\mathbf{z}_i, \mathbf{z}_j))_{1 \le i,j \le n}$ will be the input of the predictor space in DB-LM. It should be noted that the information could come directly in the form of a squared distance matrix and raw data (matrix $\mathbf{Z}$ of predictor values) may not be available, which is an advantage of distance-based prediction models over classical ones.

We define the $n \times n$ inner-products or Gram matrix as

$$\mathbf{G}_{\boldsymbol{w}} = -\frac{1}{2}\,\mathbf{J}_{\boldsymbol{w}} \cdot \mathbf{\Delta} \cdot \mathbf{J}_{\boldsymbol{w}}^\top,$$

where $\mathbf{J}_{\boldsymbol{w}} = \mathbf{I}_n - \mathbf{1} \cdot \mathbf{w}^\top$ is the $\mathbf{w}$-centering matrix, $\mathbf{I}_n$ the identity matrix of size $n \times n$, and the standardized Gram matrix as

$$\mathbf{F}_{\boldsymbol{w}} = \mathbf{D}_{\boldsymbol{w}}^{1/2} \cdot \mathbf{G}_{\boldsymbol{w}} \cdot \mathbf{D}_{\boldsymbol{w}}^{1/2}, \tag{1}$$

where $\mathbf{D}_{\boldsymbol{w}} = \mathrm{diag}(\mathbf{w})$ is a diagonal matrix whose diagonal entries are the weights in $\mathbf{w}$. A matrix $\mathbf{X}_{\boldsymbol{w}}$ with dimensions $n \times k$ is called a Euclidean configuration of $\mathbf{\Delta}$ if $\mathbf{G}_{\boldsymbol{w}} = \mathbf{X}_{\boldsymbol{w}} \cdot \mathbf{X}_{\boldsymbol{w}}^\top$, with the requirements that $rank(\mathbf{G}_{\boldsymbol{w}}) \le k \le n-1$, and matrix $\mathbf{X}_{\boldsymbol{w}}$ is $\mathbf{w}$-centered, that is, $\mathbf{w}^\top \cdot \mathbf{X}_{\boldsymbol{w}} = \mathbf{0}$. A decomposition of this kind is feasible if and only if $\mathbf{G}_{\boldsymbol{w}}$ is a positive semidefinite matrix. If this is not the case, several transformations can be applied to $\mathbf{\Delta}$ to fulfill this requirement [1]. When $\mathbf{G}_{\boldsymbol{w}}$ is positive semidefinite, $\mathbf{\Delta}$ is referred to as Euclidean (or it is said that it fulfills the Euclidean property). In classical MDS matrix $\mathbf{X}_{\boldsymbol{w}}$ is obtained through the spectral decomposition of Eq. (1), that is, given $\mathbf{F}_{\boldsymbol{w}} = \mathbf{U} \cdot \mathbf{\Lambda}^2 \cdot \mathbf{U}^\top$, where $\mathbf{\Lambda}^2$ is a diagonal matrix containing the eigenvalues of $\mathbf{F}_{\boldsymbol{w}}$, ordered in descending order, and $\mathbf{U}$ is the matrix whose columns are the corresponding eigenvectors, then $\mathbf{X}_{\boldsymbol{w}} = \mathbf{D}_{\boldsymbol{w}}^{-1/2} \cdot \mathbf{U} \cdot \mathbf{\Lambda}$. In this context, the geometric variability of $\mathbf{\Delta}$ is defined as

$$V_{\mathbf{\Delta}} = \mathrm{tr}(\mathbf{F}_{\boldsymbol{w}}) = \frac{1}{2}\mathbf{w}^\top \cdot \mathbf{\Delta} \cdot \mathbf{w}, \tag{2}$$

which serves as an extension of the concept of total variation.

2.1.2. Definition of DB-LM

For the definition of DB-LM we follow [8]. A response variable $\mathbf{Y}$, weight vector $\mathbf{w}$ and a squared distance matrix $\mathbf{\Delta}$ follow a DB-LM when the mean $\boldsymbol{\mu} = E(\mathbf{Y})$, which is $\mathbf{w}$-centered, belongs to the column space φ of $\mathbf{G}_{\boldsymbol{w}}$. It should be noted that φ is also the column space of any Euclidean configuration $\mathbf{X}_{\boldsymbol{w}}$ of $\mathbf{\Delta}$ because $\mathbf{G}_{\boldsymbol{w}} = \mathbf{X}_{\boldsymbol{w}} \cdot \mathbf{X}_{\boldsymbol{w}}^\top$.

Let $\mathbf{y}$ represent the observed values of the response variable $\mathbf{Y}$. The estimation of the DB-LM corresponding to responses $\mathbf{y}$, weights $\mathbf{w}$ and squared distances matrix $\mathbf{\Delta}$, is carried out by conducting a weighted least squares (WLS) regression of $\mathbf{y}$ on a $\mathbf{w}$-centered Euclidean configuration of $\mathbf{\Delta}$, denoted as $\mathbf{X}_{\boldsymbol{w}}$, and referred to as a latent Euclidean configuration.

Given a new case Ω_{n+1}, for which the distances to each individual in Ω are known, the new case Ω_{n+1} can be expressed as a k-vector $\mathbf{x}_{n+1}$ in the row space of $\mathbf{X}_{\boldsymbol{w}}$ using Gower's interpolation formula (see [14] and refer to [7] for the weighted version). Subsequently, the predicted response value for Ω_{n+1} is given by $\mathbf{x}_{n+1} \cdot \hat{\boldsymbol{\beta}}$, where $\hat{\boldsymbol{\beta}}$ represents the vector of estimated regression coefficients. Indeed, the DB-LM is independent of a specific $\mathbf{X}_{\boldsymbol{w}}$, as the final quantities are directly derived

from the distances. Typically, there is no necessity to explicitly define such a configuration, and the same applies to $\hat{\boldsymbol{\beta}}$ or $\mathbf{x}_{n+1}$.

In DB-LM the hat matrix is given by

$$\mathbf{H}_{\boldsymbol{w}} = \mathbf{G}_{\boldsymbol{w}} \cdot \left(\mathbf{D}_{\boldsymbol{w}}^{1/2} \cdot \mathbf{F}_{\boldsymbol{w}}^{+} \cdot \mathbf{D}_{\boldsymbol{w}}^{1/2}\right),$$

where $\mathbf{F}_{\boldsymbol{w}}^{+}$ is the Moore–Penrose pseudo-inverse of the standardized Gram matrix $\mathbf{F}_{\boldsymbol{w}}$ defined in Eq. (1). Thus, $\mathbf{H}_{\boldsymbol{w}}$ can be expressed directly as a function of the distances or, equivalently, the Gram matrix.

The DB-LM fitted values are given by

$$\hat{\mathbf{y}} = \bar{y}_{\boldsymbol{w}} \cdot \mathbf{1} + \mathbf{H}_{\boldsymbol{w}} \cdot \left(\mathbf{y} - \bar{y}_{\boldsymbol{w}} \cdot \mathbf{1}\right),$$

where $\bar{y}_{\boldsymbol{w}} = \mathbf{w}^{\top} \cdot \mathbf{y}$ is the $\mathbf{w}$-mean of $\mathbf{y}$.

The predicted response value for a new case Ω_{n+1} is given by

$$\hat{y}_{n+1} = \bar{y}_{\boldsymbol{w}} + \frac{1}{2}\left(\mathbf{g}_{\boldsymbol{w}} - \boldsymbol{\delta}_{n+1}\right) \cdot \left(\mathbf{D}_{\boldsymbol{w}}^{1/2} \cdot \mathbf{F}_{\boldsymbol{w}}^{+} \cdot \mathbf{D}_{\boldsymbol{w}}^{1/2}\right) \cdot \left(\mathbf{y} - \bar{y}_{\boldsymbol{w}} \cdot \mathbf{1}\right),$$

where $\mathbf{g}_{\boldsymbol{w}}$ denotes the $1 \times n$ row vector containing the necessarily non-negative diagonal entries of $\mathbf{G}_{\boldsymbol{w}}$ and $\boldsymbol{\delta}_{n+1}$ is the $1 \times n$ row vector of squared distances from Ω_{n+1} to each individual in Ω.

DB-LM encompasses WLS as a specific case: if we begin with an $n \times r$ $\mathbf{w}$-centered matrix $\mathbf{X}_{\boldsymbol{w}}$ containing r continuous predictors for n individuals, and define $\boldsymbol{\Delta}$ as the matrix of squared Euclidean distances between the rows of $\mathbf{X}_{\boldsymbol{w}}$, then $\mathbf{X}_{\boldsymbol{w}}$ is trivially a Euclidean configuration of $\boldsymbol{\Delta}$. Consequently, the DB-LM hat matrix, response, and predictions align with the corresponding WLS quantities of an ordinary linear model.

2.2. Predictor distance matrices

Let $\mathbf{z}_i, \mathbf{z}_j$ be two rows of the $n \times p$ matrix $\mathbf{Z}$ corresponding to the measurements of the predictor variables for individuals Ω_i, Ω_j, $i, j = 1, \ldots, n$. A well-known and commonly used distance for mixed type data is Gower's distance, which was defined in [11] as

$$\delta(\mathbf{z}_i, \mathbf{z}_j) = 1 - \frac{\sum_{h=1}^{p_1}\left(1 - \left|z_{ih} - z_{jh}\right|/G_h\right) + a + \alpha}{p_1 + \left(p_2 - d\right) + p_3},$$

with p_1 the number of continuous variables, G_h the range of the hth continuous variable, a and d the number of positive and negative matches, respectively, for the p_2 binary variables, and α the number of matches for the p_3 multi-state categorical variables. Note that the total number of variables is $p = p_1 + p_2 + p_3$.

Gower's distance can be defined as the Pythagorean sum of three distance measures for quantitative, binary and multi-state categorical variables, $\boldsymbol{\Delta}_1 = \delta_1^2(\mathbf{z}_i, \mathbf{z}_j))$, $\boldsymbol{\Delta}_2 = \delta_2^2(\mathbf{z}_i, \mathbf{z}_j))$ and $\boldsymbol{\Delta}_3 = \delta_3^2(\mathbf{z}_i, \mathbf{z}_j))$ for $1 \le i, j \le n$, where δ_1 is the range-normalized city-block distance, δ_2 distance is associated to Jaccard's similarity coefficient and δ_3 is the Hamming distance. However, this classical distance presents two main drawbacks: It does not take into account the correlations between quantitative variables and it is a not robust metric.

In [12,15] two robust alternatives to Gower's distance were proposed in the context of MDS and clustering for the non-weighted case. In the latter, authors studied a robustification of Gower's distance by taking δ_1 as a robust Mahalanobis distance, instead of the range-normalized city-block one, and δ_2 and δ_3 were left unchanged (all of them conveniently standardized to equal geometric variability in order to be commensurate). In the former, these three distances were combined via Related Metric Scaling (RelMS) [16] to obtain a joint metric able to discard redundant information coming from different sources. We explicit here the formulation in the weighted context (see [17] for an extension to $m > 3$ sources). Let $\boldsymbol{\Delta}_l$ for $l = 1, 2, 3$, be three matrices of squared distances, each corresponding to a different variable type (considered here as different sources of information), with equal geometric variability. For each $\boldsymbol{\Delta}_l$, consider its Gram matrix $\mathbf{G}_{\boldsymbol{w},l}$ and its standardized version $\mathbf{F}_{\boldsymbol{w},l} = \mathbf{D}_{\boldsymbol{w}}^{1/2} \cdot \mathbf{G}_{\boldsymbol{w},l} \cdot \mathbf{D}_{\boldsymbol{w}}^{1/2}$. Then the joint metric is obtained by combining the corresponding standardized Gram matrices as follows:

$$\mathbf{F}_{\boldsymbol{w}}^{J} = \sum_{l=1}^{3} \mathbf{F}_{\boldsymbol{w},l} - \frac{1}{3}\sum_{l \neq m} \mathbf{F}_{\boldsymbol{w},l}^{1/2} \cdot \mathbf{F}_{\boldsymbol{w},m}^{1/2}, \tag{3}$$

where $\mathbf{F}_{\boldsymbol{w},l}^{1/2}$ denotes the square root of $\mathbf{F}_{\boldsymbol{w},l}$ for $l = 1, 2, 3$. The final RelMS metric $\boldsymbol{\Delta}$ can be recovered from (3) as follows:

$$\boldsymbol{\Delta} = \mathbf{1} \cdot \mathbf{g}_{\boldsymbol{w}}^{J} + (\mathbf{g}_{\boldsymbol{w}}^{J})^{T} \cdot \mathbf{1}^{T} - 2\,\mathbf{G}_{\boldsymbol{w}}^{J},$$

where $\mathbf{G}_{\boldsymbol{w}}^{J} = \mathbf{D}_{\boldsymbol{w}}^{-1/2} \cdot \mathbf{F}_{\boldsymbol{w}}^{J} \cdot \mathbf{D}_{\boldsymbol{w}}^{-1/2}$ and $\mathbf{g}_{\boldsymbol{w}}^{J}$ is a row vector containing the diagonal of matrix $\mathbf{G}_{\boldsymbol{w}}^{J}$. Note that the first addend of (3) mimics Gower's distance by adding the three metrics (through the standardized Gram matrices, in this case), whereas the second one is responsible of discarding redundant information coming from different sources (see [12,17,18]).

In this paper we propose a new robust distance for quantitative data, either when using robust Gower's distance or robust RelMS. Thus, we propose to take δ_1 as a new robust Mahalanobis distance, computed in terms of a robust covariance matrix estimated from a distance-based trimming estimator.

The distance-based trimming estimator is defined in terms of a proximity function used in [12] to detect outliers in complex datasets. In particular, consider $\mathbf{z}_1, \ldots, \mathbf{z}_n$ the rows of the $n \times p$ matrix $\mathbf{Z}$ of predictor measurements and $\boldsymbol{\Delta} = (\delta^2(\mathbf{z}_i, \mathbf{z}_j))_{1 \le i,j \le n}$ the corresponding matrix of squared pairwise distances. Given a new individual Ω_0 with measurements $\mathbf{z}_0 \in \mathbb{R}^p$, the proximity function of $\mathbf{z}_0$ to the set $\{\mathbf{z}_1, \ldots, \mathbf{z}_n\}$ is defined, in the weighted context, as

$$\phi(\mathbf{z}_0) = \sum_{i=1}^{n} w_i \cdot \delta^2(\mathbf{z}_0, \mathbf{z}_i) - V_{\boldsymbol{\Delta}}, \tag{4}$$

where w_i is the weight for individual Ω_i and $V_{\boldsymbol{\Delta}}$ is the geometric variability of $\boldsymbol{\Delta}$ defined in (2).

The procedure to obtain a robust estimation of the covariance matrix for the quantitative predictors is as follows: Let $\mathbf{Z}^{p_1}$ be the $n \times p_1$ matrix of continuous predictors for the n individuals. We start by $\mathbf{w}$-centering it, that is, $\mathbf{Z}_{\boldsymbol{w}}^{p_1} = \mathbf{D}_{\boldsymbol{w}}^{1/2} \cdot \mathbf{J}_{\boldsymbol{w}} \cdot \mathbf{Z}^{p_1}$. Next, we proceed by calculating the matrix of pairwise squared Mahalanobis distances between the rows of $\mathbf{Z}_{\boldsymbol{w}}^{p_1}$, and denote it by $\boldsymbol{\Delta}_{mah}$. For each unit in the dataset we use function (4) to compute its proximity to the remaining $n-1$ units taking into account the entries in $\boldsymbol{\Delta}_{mah}$ and the corresponding geometric variability. Next, units are sorted in ascending order according to their ϕ values. The greater the value of ϕ the further the unit from the bulk of the data. Finally, the trimming is performed by excluding a given percentage of the data, say α, according to their ϕ values. We obtain the α-trimmed sample $\mathbf{Z}_{\alpha}^{p_1}$ and the associated $\mathbf{J}_{\boldsymbol{w},\alpha}$ and $\mathbf{D}_{\boldsymbol{w},\alpha}$, from which we calculate $\mathbf{Z}_{\boldsymbol{w},\alpha}^{p_1} = \mathbf{D}_{\boldsymbol{w},\alpha}^{1/2} \cdot \mathbf{J}_{\boldsymbol{w},\alpha} \cdot \mathbf{Z}_{\alpha}^{p_1}$ the $\mathbf{w}$-centered α-trimmed sample. Finally, we calculate a distance-based α-trimmed estimation of the covariance matrix as

$$\hat{\mathbf{S}}_{\alpha} = \mathbf{Z}_{\boldsymbol{w},\alpha}^{p_1\,\top} \cdot \mathbf{Z}_{\boldsymbol{w},\alpha}^{p_1} = \mathbf{Z}_{\alpha}^{p_1\,\top} \cdot \mathbf{J}_{\boldsymbol{w},\alpha}^{\top} \cdot \mathbf{D}_{\boldsymbol{w},\alpha} \cdot \mathbf{J}_{\boldsymbol{w},\alpha} \cdot \mathbf{Z}_{\alpha}^{p_1},$$

and robust pairwise Mahalanobis distances are obtained accordingly, i.e.,

$$\delta^2(\mathbf{z}_i^{p1}, \mathbf{z}_j^{p1}) = (\mathbf{z}_i^{p1} - \mathbf{z}_j^{p1})^{\top} \cdot \hat{\mathbf{S}}_{\alpha}^{-1} \cdot (\mathbf{z}_i^{p1} - \mathbf{z}_j^{p1}),$$

where $\mathbf{z}_i^{p1}$ denotes the quantitative measurements for individual Ω_i.

3. Simulation study

3.1. Scenarios considered

The performance of the new proposals, robust G-Gower's and robust RelMS, is evaluated and compared to those of Euclidean, classical Gower's and G-Gower's metrics in several scenarios with a given percentage of outlier contamination.

In particular, three datasets of size $n = 500$ were generated with $p_1 = 4$ quantitative predictors, $p_2 = 2$ binary predictors and $p_3 = 3$ multi-state categorical ones, with different correlation/association structure among predictors. Next, outliers were introduced by perturbing several values and/or characteristics in the explanatory variables of existing units as follows. For quantitative predictors fluctuations of 3 times the corresponding SD were added to perturbed units, whereas for categorical predictors unit characteristics were changed in a contradictory way. As a result, twelve scenarios were considered:

1. Highly correlated/associated predictors with a 10% outlier contamination in
   (a) All quantitative predictors.
   (b) All multi-state categorical predictors.
   (c) All binary predictors.
   (d) All predictors.
2. Intermediate correlated/associated predictors with a 5% outlier contamination in
   (a) All quantitative predictors.
   (b) All multi-state categorical predictors.
   (c) All binary predictors.
   (d) All predictors.
3. Intermediate correlated/associated predictors with a 5% outlier contamination in
   (a) Only one quantitative predictor.
   (b) Only one multi-state categorical predictor,
   (c) Only one binary predictor.
   (d) One predictor of each type.

For each scenario, two types of response variables were obtained, either as a random linear or non-linear combination of the predictors. Additionally, models performance was evaluated on contaminated and uncontaminated responses, giving rise to four kinds of response (linear/non-linear, contaminated/uncontaminated) for each considered scenario.

The linear response was generated from the sum of the main effects of a linear model with coefficients equal to 1 plus a standardized normal random error. The information of the categorical predictors was included taking into account the binary variables of each class as usual, including as many binary variables as classes minus one for each predictor. In the case of linear response, the underlying model would be the one that would correspond to using Euclidean distance, since the predictors themselves could already form a possible configuration of latent variables. The response that we call non-linear was generated from the sum of the main effects of all predictors and with the subtraction of second order interactions between the three sets of variable types, that is, the sets of quantitative, categorical and binary variables. The idea in the non-linear case was to generate a response variable that included the information of the main effects excluding the first order dependencies between the sets of variables. All of them with coefficients equal to 1 plus a standardized normal random error. The non-linear model would correspond to a case where the Gower metric would, in principle, improve the fit compared to the Euclidean one. Gower tends to have more dimensions than the Euclidean case and fits non-linear situations better (see e.g. [6,8]). In this paper the idea is to compare the fit of the proposed metrics, robust G-Gower and robust RelMS with those of the classical ones, like Euclidean and Gower's, as well as with G-Gower's.

The contaminated response was generated from the contaminated predictors. The uncontaminated response was obtained by leaving unchanged the response variable of those units whose predictor's values were perturbed.

Fig. 1 contains the MDS configurations of the synthetic datasets regarding scenarios (2a): Intermediate correlated/associated predictors with a 5% outlier contamination in all quantitative predictors and (3d): Intermediate correlated/associated predictors with a 5% outlier contamination in one predictor of each type, with the aim of illustrating the metrics' behavior in the presence of outliers. Diagonal panels contain conditional histograms regarding data type (outlier or not) and off-diagonal panels show the corresponding MDS maps. These multiple scatter plots were produced with Matlab's function `gplotmatrix`. Classical Gower's metric is compared to two robust proposals, such as robust G-Gower, where a robust Mahalanobis distance was used for quantitative predictors instead of range-normalized city block distance, and robust RelMS, computed from formula (3) as a combination of robust Mahalanobis for quantitative predictors, distance associated to Jaccard's similarity coefficient for binary predictors and Hamming distance for multi-state categorical ones. Robust proposals were estimated by means of the distance-based trimming estimator defined in (4) with a 5% trimming threshold. In general, we observe that most of the outliers are placed apart from the bulk of the data when using any of the proposed robust metrics. This is the reason why we propose to built the DB-LM using a robust metric in the predictors' space. In the next section we see that these models are more effective in the prediction of the response variable when data contain anomalous units.

3.2. Results

The effectiveness in the prediction of the response was evaluated through leave-one-out. For each scenario nine predictive models based on different metrics were fitted using the `dblm` function of the `dbstats` package for R. See Appendix A for the usage of the function. Next, squared errors (SEs) were computed including its mean, median and standard deviations (SDs). In all cases an explanation of 80% of the geometric variability was required by setting `rel.gvar` parameter equal to 0.8 in the `dblm` function. In this way no model was overparameterized and, at the same time, sufficient information from the predictor space was included. Indeed, the selection of the number of latent dimensions can be done automatically through argument "method" in `dblm` and `dbglm` functions (we refer to the package help in cran for the detail). Method can be equal to

- `eff.rank`: the user can choose the effective rank equal to a fixed number of latent variables.
- `rel.gvar`: the user can choose a fixed percentage of geometric variability corresponding to a given number of latent dimensions, depending on the coordinates of the predictor space in each real dataset.
- `OCV, GCV, AIC or BIC`: depending on the function used in `dbstats`, it is possible to choose one of the following criteria (ordinary cross-validation, generalized cross-validation, Akaike or Bayesian information criteria) to select the number of dimensions optimally. In addition, it is possible to plot the results of the selected statistic using the plot command by specifying the argument "which = c(6)".

Since the aim of the paper is to study the use of robust metrics that depend on a trimming parameter in the big data context, in this simulation study the percentage of explained variability was set to 80%. Other methods for latent variable selection are explored in Section 4. Regarding the robust proposals, we decided to consider a range of trimming thresholds in order to cover the true outlier percentage in each considered scenario and explore accordingly the models' performance. Metrics under evaluation were:

- Euclidean,
- classical Gower's,
- Generalized Gower's (G-Gower), where range-normalized city-block distance was substituted by Mahalanobis distance,

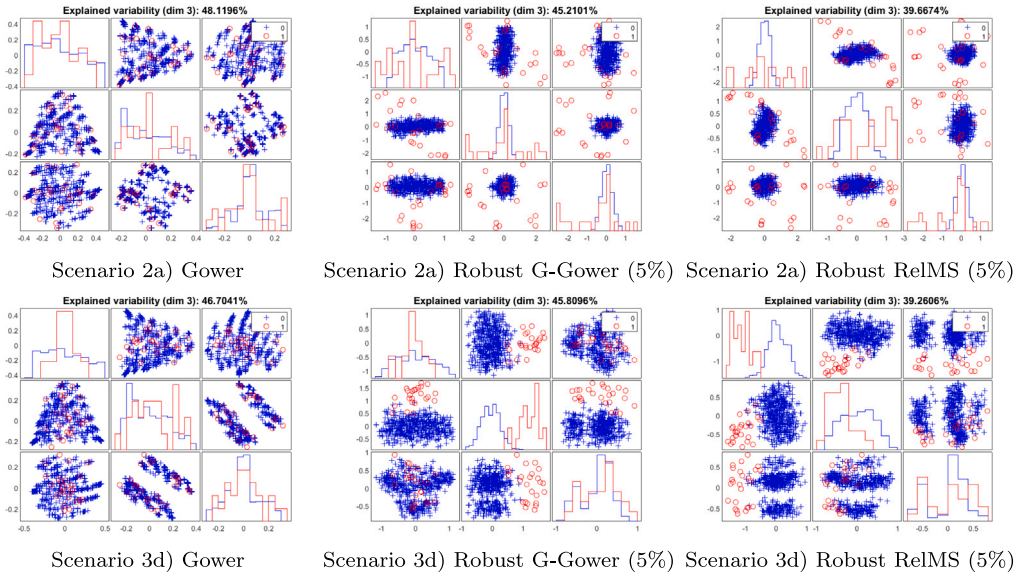

Fig. 1. MDS configurations obtained from Gower's, robust G-Gower and robust RelMS metrics for Scenarios (2a) and (3d). Outliers are depicted in red color.

- Robust generalized Gower (robust G-Gower), where robust Mahalanobis distance was used for quantitative predictors, estimated by means of the distance-based trimming estimator defined in (4) with trimming thresholds of 5%, 10% and 15%.
- Robust RelMS, computed from formula (3) as a combination of robust Mahalanobis for quantitative predictors, distance associated to Jaccard's similarity coefficient for binary predictors and Hamming distance for multi-state categorical ones. As before, the distance-based trimming estimator defined in (4) with trimming thresholds of 5%, 10% and 15% was used to estimate Mahalanobis distance.

As an illustration of models' performance, from Figs. 2 to 6 we depict the box-plots of the SE distributions concerning DB-LM based on Gower's, robust G-Gower's (with 5% and 10% trimming) and robust RelMS (with 5% and 10% trimming) for all scenarios considered. In particular, Fig. 2 contains the models' performance on uncontaminated datasets, where we can already observe as general conclusions that in the case of linear response Gower fits the model worse than robust G-Gower and robust RelMS, as expected. In all cases, the best fit is attained by robust RelMS model, observing that when there is an intermediate correlation/association the robust RelMS model outperforms robust G-Gower for linear and nonlinear responses. In the case of highly correlated/associated predictors both metrics provide an equally good fit to the scenario.

Models' performance on contaminated datasets can be found from Figs. 3 to 6. Figures' panels are organized as follows: Scenarios 1–3 corresponding to three datasets with different correlation/association structure are placed by columns, and cases (a)–(d), which indicate which predictors were contaminated, are placed by rows. Note that, the scale of the vertical axis is the same within figures, for better comparison.

Additionally, summary statistics for the estimated SE values can be found in Appendix B (see from Tables 6 to 13).

In Fig. 3 we can observe that when the response variable is an uncontaminated random linear combination of the predictors the best performance is obtained by DB-LM with robust RelMS metrics in nine out of twelve scenarios, and by robust G-Gower's in two out of the three remaining cases. We reach to similar conclusions by looking at SE mean and median values shown in Table 6 in Appendix B.

In the case of linear and contaminated response (see Fig. 4) the DB-LM with robust RelMS metrics has the best performance in nine out of twelve scenarios and robust G-Gower's in the remaining three cases. These results can also be observed in Tables 7 and 11 in Appendix B, where robust RelMS metrics attain the lowest SE mean, median and SD values in these scenarios.

Fig. 5 contains the SE distributions regarding the uncontaminated non-linear response, where we see that the DB-LM with robust RelMS metrics has the best performance in ten out of twelve scenarios, and classical Gower's in the remaining two cases. These results can also be observed in Table 8 and to some extent in Table 12 in Appendix B, where robust RelMS metrics attain the lowest SE mean, median and SD values in these scenarios.

Finally, for the case of contaminated non-linear response (see Fig. 6) the DB-LM with robust RelMS metrics has the best performance in nine out of twelve scenarios and classical Gower's in the remaining three cases. These results can also be observed in Table 9 and to some extent in Table 13 in Appendix B, where robust RelMS metrics attain the lowest SE mean, median and SD values in these scenarios.

Thus, the DB-LM based on a robust proposal outperforms the other studied metrics in 87.5% of the cases (77.1% robust RelMS and 10.4% robust G-Gower).

4. Application

In this section DB-LMs with different metrics are applied to two real datasets. A common feature of both real datasets is that they contain outliers and include predictors that are correlated/associated with each other, as usually happens in regression model predictor sets. Like in the simulation study, models' performance in the prediction of the response is evaluated in terms of the MSE and some summary statistics are provided. Additionally, several trimming thresholds are studied and box-plots with SE distributions are given to illustrate the results. The trimming threshold was selected within a wide range of values as that which produced the lowest MSE. One step further would be to monitor the MSE, or other statistic of interest, in order to make a data-driven selection (see [19]).

For the comparison of metrics, either a fixed percentage of variability or a fixed number of dimensions was determined. The former was used in bike-sharing demand application, whereas the latter was explored in the motor insurance dataset. The proposed Robust DB-LMs are compared to two benchmark models, for the one hand, the classical LM which appears as particular case of DB-LM with Euclidean distance and, on the other hand, DB-LM with Gower's distance which has been

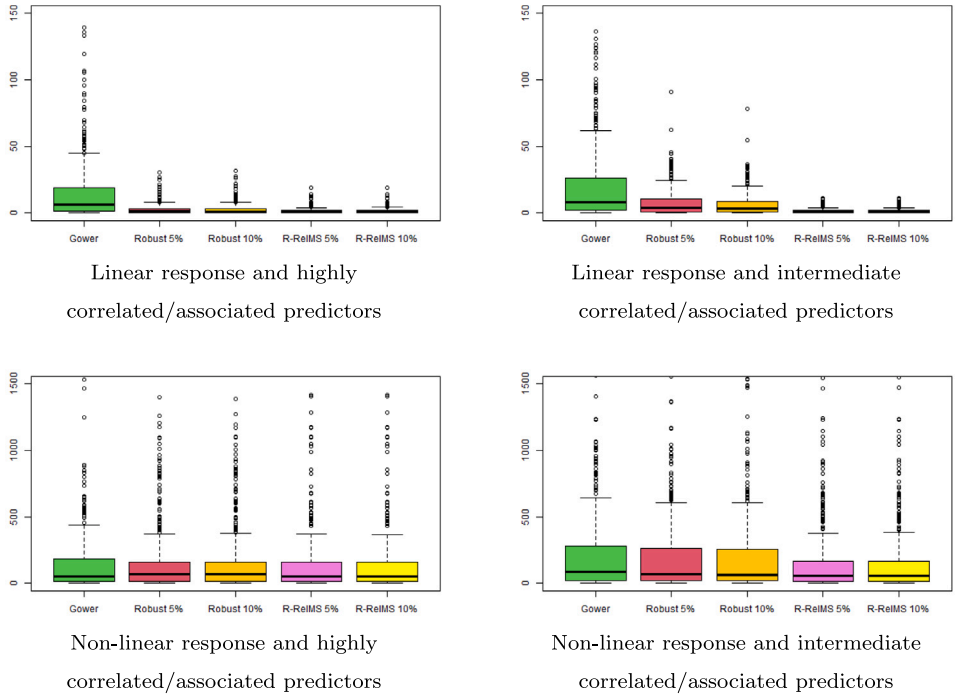

Fig. 2. SE distributions for DB-LM with Gower's, robust G-Gower (5%, 10% trimmed) and robust RelMS (5%, 10% trimmed) metrics. Uncontaminated datasets.

traditionally used when dealing with mixed type data. In all cases, Robust DB-LMs outperform benchmark models registering lower MSE. We leave for further research the treatment of both trimming threshold and latent dimension selection in the context of big data.

4.1. Bike sharing demand

This dataset was prepared and firstly analyzed by [10]. Data comes from two different sources. Bike demand information is provided by the Capital Bikeshare program and it is available at their website (www.capitalbikeshare.com/system-data, accessed on 3 June 2019). Capital Bikeshare operates in the District of Columbia, Arlington County, VA, and the City of Alexandria, VA. The program records several details, such as travel duration, departure and arrival locations, and time elapsed between departure and arrival, for each rental in the bike sharing system. Following [10] we decided to analyze the data on a daily basis (hourly information was also available) for the period between 1 January 2013 and 31 December 2018. The scarce days for which data were not available were deleted from the sample. The variables appearing in the Capital Bikeshare files for each day are listed in the first part of Table 1. A second source of information regarding weather conditions was added to the data provided by Capital Bikeshare (year, month, day, and count of users). Thus, for each day, variables describing weather conditions that could affect the decision of picking up or not a bike (second part of Table 1). This information at Ronald Reagan Washington National Airport (DCA) was gathered from the website of the National Oceanic and Atmospheric Administration (NOAA). The reason for choosing the specific DCA location was to ensure data availability for all days of interest and also because it is centered in the area covered by the bike stations, and thus it is a good representative. The NOAA variables in the second part of Table 1 are treated as quantitative ones. The variables in the first part of Table 1 are treated as qualitative ones (declared as factors in R), except for the total count of daily users. The response variable is the daily count of users (casual and registered) scaled by the mean daily number of users (in the year corresponding to the day). The total number of days in the dataset is 2903.

Table 1
Variables included in Bike sharing dataset.

Capital Bikeshare
Total count of daily users (both registered and not)
Season: winter (1), spring (2), summer (3), autumn (4)
Year, codified to 0 (= 2011), 1 (= 2012), 2 (= 2013), ..., 7 (= 2018)
Month, codified to 1, 2, ..., 12
National holiday (1) or not (0)
Weekday, codified to 0 (= Sunday), 1 (= Monday), ..., 6 (= Saturday)
Working day (1) or weekend day (0)
NOAA at DCA
Average daily wind speed (miles per hour)
Precipitation (inches to hundredths)
Maximum temperature (in Fahrenheit)
Minimum temperature (in Fahrenheit)
Ceiling height dimension (in meters)
Mean daily temperature (in Celsius)
Sea level pressure (in hPa)
Relative humidity (in %)

Cross-validation was used to evaluate the effectiveness in the prediction of responses. Thus, DB-LM with eleven different metrics was estimated with a training sample of 70% of the data and tested in the remaining 30%, where SE was computed from 300 runs. Despite the high dimensionality of computations, cross-validation was run a high number of times to ensure the outlier selection within samples. All models were fitted with `rel.gvar` = 0.5. This is because Gower had 2902 latent variables (in this case $n-1$) if all the geometric variability was included, i.e., `rel.gvar` = 1. Additionally, box-plots of the SEs were produced (see Fig. 7).

Several trimming thresholds were considered when studying the performance of the DB-LM with robust metrics. In particular, 3%, 5%, 10% and 15% thresholds were analyzed in order to study the behavior of the MSE. From Fig. 7 and Table 2 we observe that for robust G-Gower the lowest MSEs is attained at 3% and 5% trimming, and at 3% trimming for RelMS. The model with the lowest MSE among those that were considered is robust RelMS with a 3% trimming. A summary of the output for the two best models can be found in Appendix C.

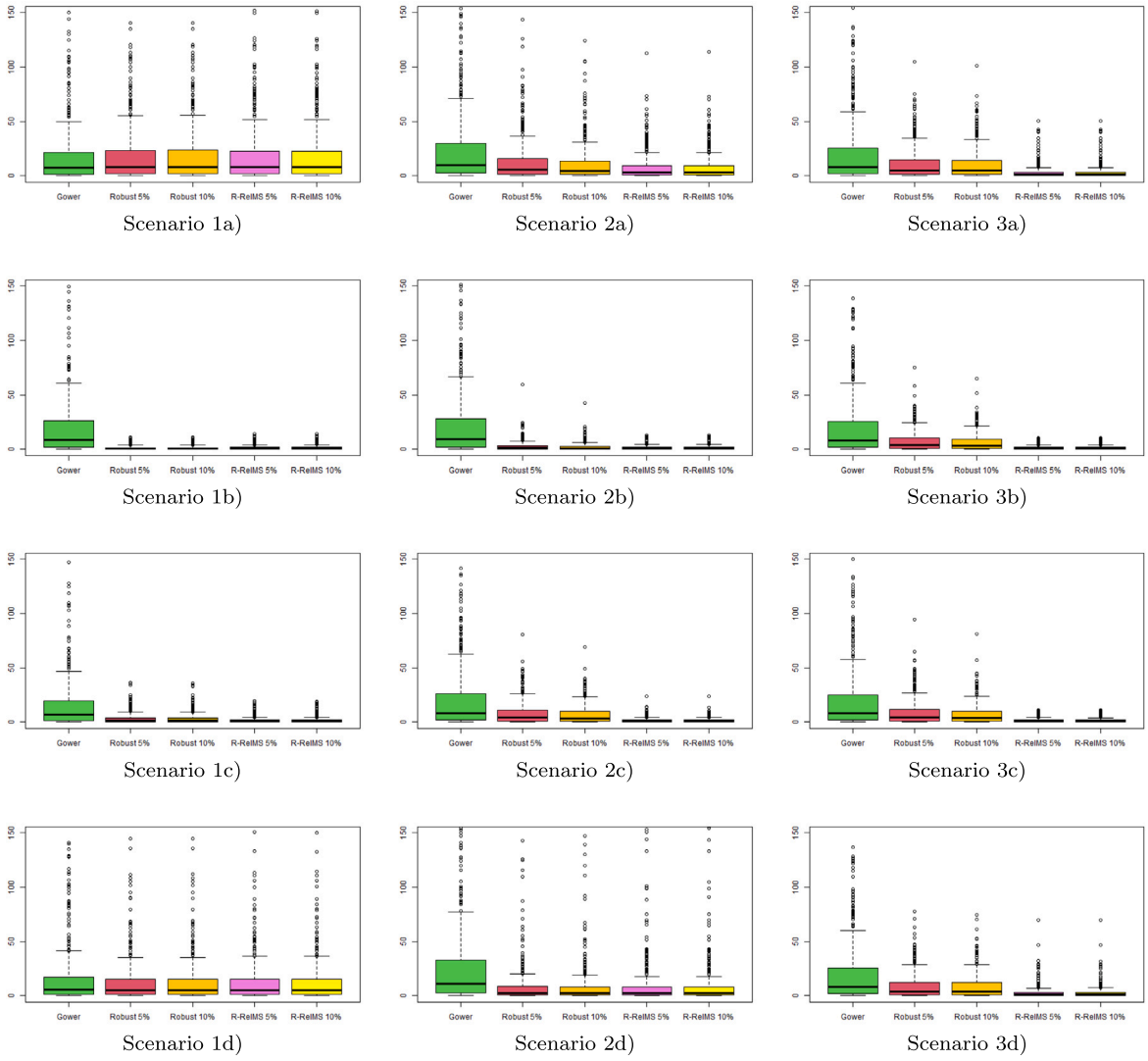

Fig. 3. SE distributions for DB-LM with Gower's, robust G-Gower (5%, 10% trimmed) and robust RelMS (5%, 10% trimmed) metrics, scenarios (1a) to (3d). Uncontaminated linear response.

Table 2
Estimated SE (mean, median and SD values) for the DB-LM with different metrics. Bike sharing dataset.

Metric	Mean	Median	SD
Euclidean	0.1334	0.1331	0.0038
Gower	0.0410	0.0412	0.0022
G-Gower	0.0366	0.0365	0.0018
Robust G-Gower (3%)	0.0361	0.0361	0.0016
Robust G-Gower (5%)	0.0355	0.0355	0.0016
Robust G-Gower (10%)	0.0411	0.0408	0.0024
Robust G-Gower (15%)	0.0414	0.0410	0.0024
Robust RelMS (3%)	0.0347	0.0347	0.0019
Robust RelMS (5%)	0.0350	0.0350	0.0019
Robust RelMS (10%)	0.0359	0.0359	0.0019
Robust RelMS (15%)	0.0365	0.0365	0.0020

4.2. Motor insurance dataset

We fitted DB-LMs to data on Swedish third-party motor insurance in 1977 from [20]. Data can be downloaded from `faraway` R package named `motorins` [21]. The data come from a study conducted by a committee on risk premiums in automobile insurance. A subset of this data, specifically the records with Zone = 1, corresponds to Stockholm, Göteburg, and Malmo. The total number of observations (for Zone = 1) is $n = 295$. The recorded variables for each risk group include *Payment* (total payments in Skr), *Claims* (number of claims), and *Insured* (number of insured, in policy-years). These data could be utilized to exemplify premium rating, wherein risk premiums are computed as the product of claim frequency multiplied by claim severity.

Three predictors are included: *Distance* (Kilometers traveled), *Bonus* (No-claims bonus) and *Make* (specified car makes). *Distance* and *Bonus*

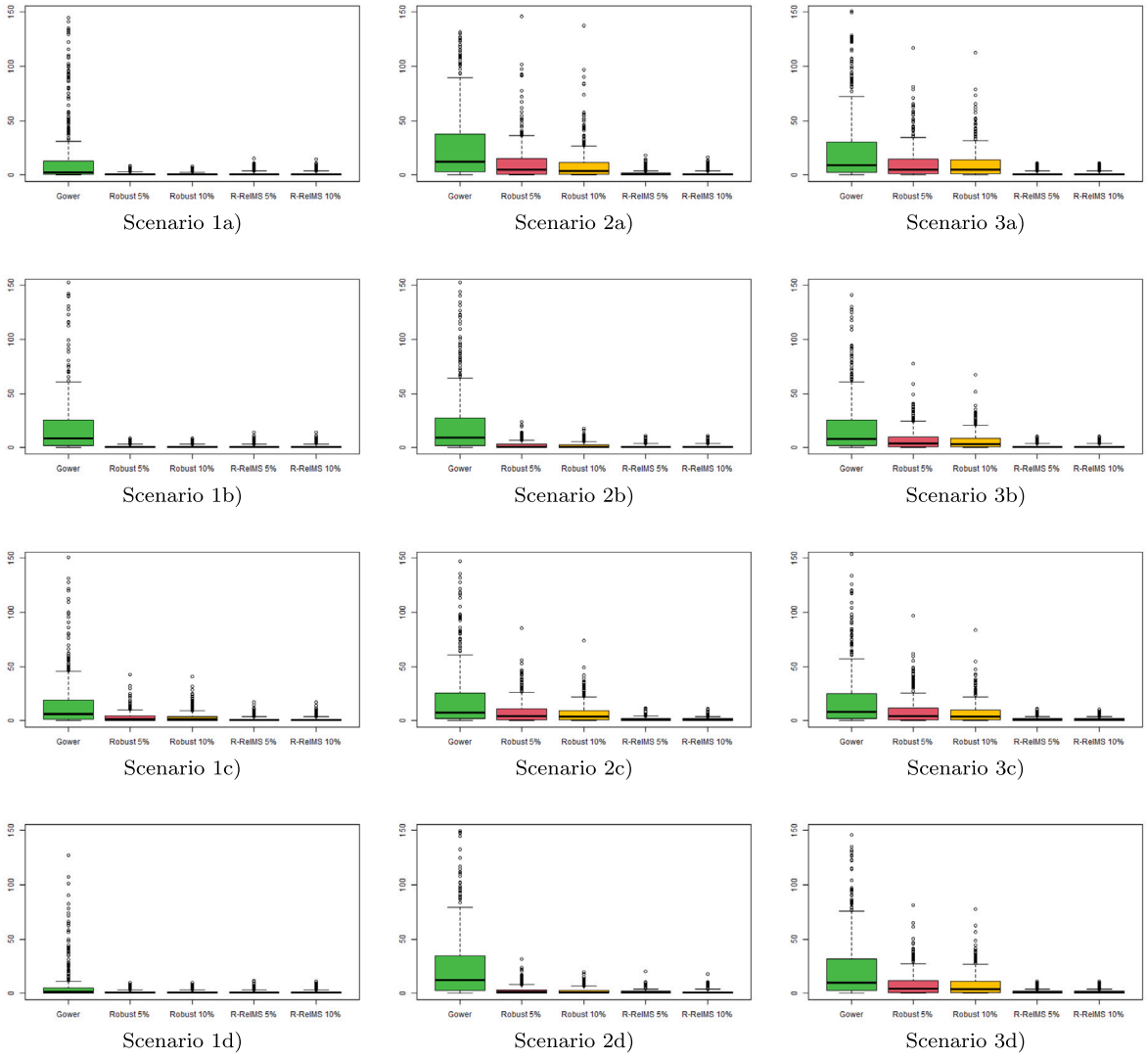

Fig. 4. SE distributions for DB-LM with Gower's, robust G-Gower (5%, 10% trimmed) and robust RelMS (5%, 10% trimmed) metrics, scenarios (1a) to (3d). Contaminated linear response.

are considered quantitative variables, while *Make* is treated as a categorical one with nine specified car makes, as performed in [8]. The two continuous predictors, *Distance* and *Bonus*, have high values in the covariance matrix. This fact suggests that the use of metrics that take into account a normalization with respect to the covariance matrix may be appropriate and may model the data properly.

This dataset was analyzed in [8], where claim severity was modeled using distance-based generalized linear models with Gamma error structure and logarithmic link and, as a result, Gower's metric outperformed the Euclidean one. Here we analyzed claim severity too, calculated as $Payment/Claims$ with weights given by the corresponding number of claims (*Claims*) by using the linear model with Gaussian error and identity link, with the aim of illustrating the performance of the DB-LM with robust metrics. The analysis was performed for different dimensions $k = 3, 4, 5, 6$ (latent variables) in order to select the dimension for each model which minimizes the corresponding MSE. The parameter in the `dblm` function to be set was `eff.rank` $= k$. In the case of robust proposals, several trimming thresholds were considered, such as 3%, 5%, 10% and 15%. For each model, SEs were estimated by leave-one-out.

As can be seen in Table 3 the lowest MSE is obtained for all the metrics when $k = 3$ latent variables are used in the DB-LM. Table 4 contains the explained percentage of geometric variability corresponding to the effective rank of $k = 3$. Given $k = 3$, the lowest MSE values are attained for the DB-LM using a robust G-Gower metric with 3% trimming and the DB-LM with G-Gower metric. Fig. 8 contains the boxplots of the SE distributions for several metrics, where it is observed, as already obtained in [8], that the classical model (when using Euclidean distance) provides a worse fit than the DB-LMs. A summary of the two best models is included in Appendix C.

The selection of the number of latent variables based on GVC, AIC and BIC criteria is illustrated with this dataset. Results are shown in Table 5 for four selected models: Robust G-Gower and Robust RelMS metrics for 3% and 5% trimming using $k = 3, 4, 5, 6$. In all cases, the minimum values are attained for $k = 3$ latent variables. Fig. 9 contains the graphical output of `dbstats` for the optimal BIC for a range of dimensions from $k = 3, 4, 5, 6$, for the Robust DB model with G-Gower's with 3% trimming when using method = "BIC". The optimal BIC value $k = 3$, colored in red, is that internally used for the model fitting.

5. Conclusions

In this paper new metrics were proposed to robustify the DB-LM, by means of a distance-based trimming statistic, in the case of

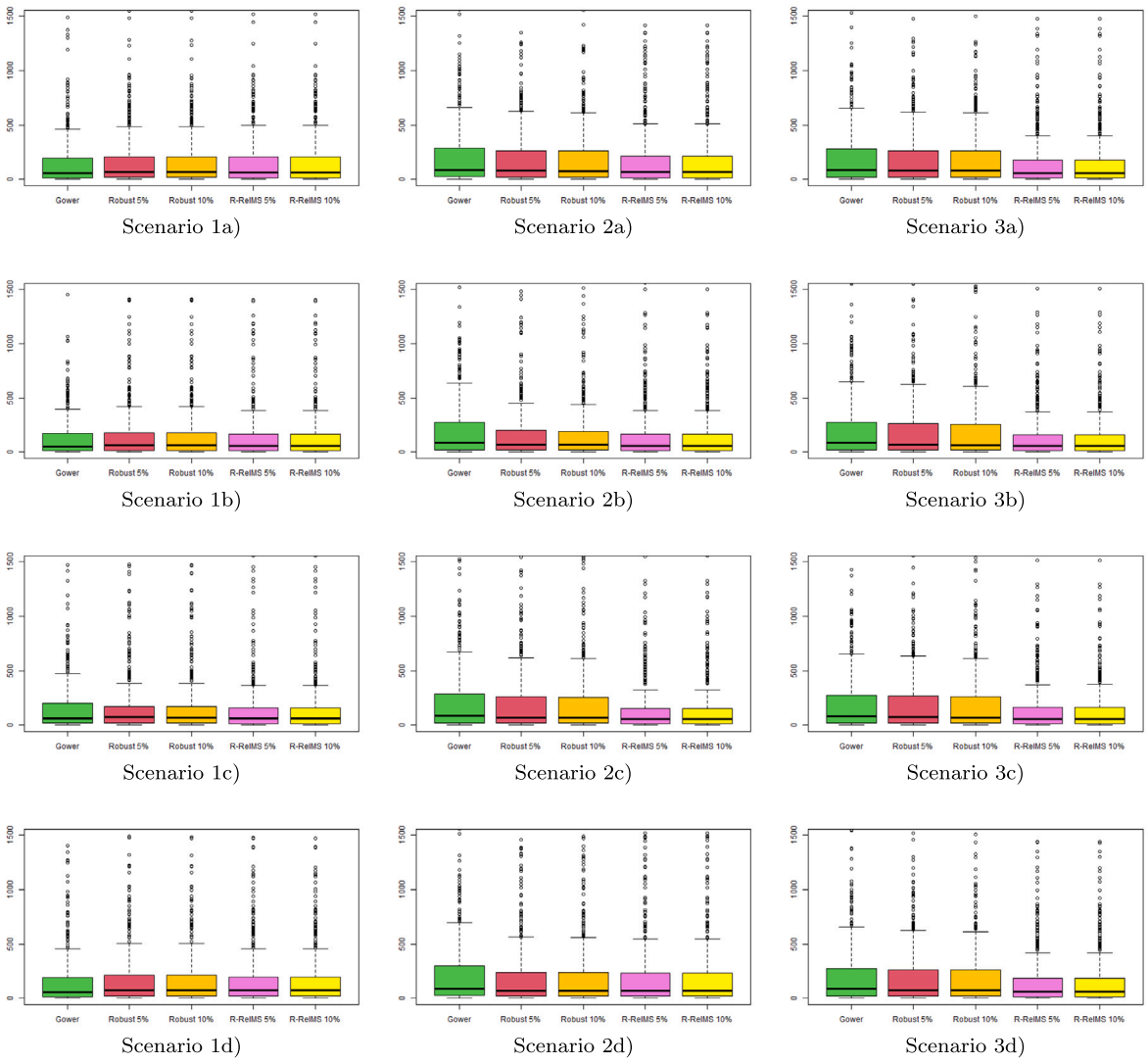

Fig. 5. SE distributions for DB-LM with Gower's, robust G-Gower (5%, 10% trimmed) and robust RelMS (5%, 10% trimmed) metrics, scenarios (1a) to (3d). Uncontaminated non-linear response.

weighted and multivariate heterogeneous predictors. In particular, the proposed metrics were robust G-Gower's and robust RelMS, with different trimming thresholds. The new metrics consider outliers, variable redundancy and provide implicit variable selection.

The performance of robust DB-LM, i.e., DB-LM using these new robust metrics, was evaluated in the presence of outliers through a simulation study involving 48 different scenarios and two real datasets, and compared to those of classical metrics, such as Euclidean and Gower's, as well as G-Gower's. For each scenario, SEs were computed via leave-one-out and MSE, median SE and SD of SE were used to evaluate the effectiveness in the prediction of responses. Box-plots with the SE distributions were provided for each scenario and several metrics. In general, we observed that, first, the DB-LM based on robust proposals outperformed those based on other metrics, such as Euclidean, Gower's or G-Gower's; second, in the case of linear response, DB-LM with Gower's metric exhibited a worse fit than robust G-Gower and robust RelMS; third, for predictors with an intermediate correlation/association structure DB-LM with robust RelMS outperformed robust G-Gower for linear and nonlinear responses and, finally, in the case of highly correlation/association among predictors robust metrics provided a similar fit to the scenario.

Additionally, the DB-LM was illustrated on two real data sets, both containing outliers and including predictors with a correlation/association structure. The first application, was framed in the area of sustainable transport, and focused on the prediction of bike sharing demand from a mixed-type set of predictors related to renting details and weather conditions. The second application was in the area of finance and banking, where the aim was to predict claim severity in motor insurance regarding a set of weighted mixed-type predictors. Models' performance was evaluated by cross-validation, and in both cases we concluded that DB-LM with robust proposals outperformed the fit of other considered metrics.

The distance-based trimming statistic was defined in a general setting, and we applied it to get a robust estimation of the covariance matrix of quantitative variables using Mahalanobis distance. Other metrics can be considered depending on the nature of the data. A deeper study is left for further research. Another interesting direction for future research is the robustification of DB-GLM, which is a generalization

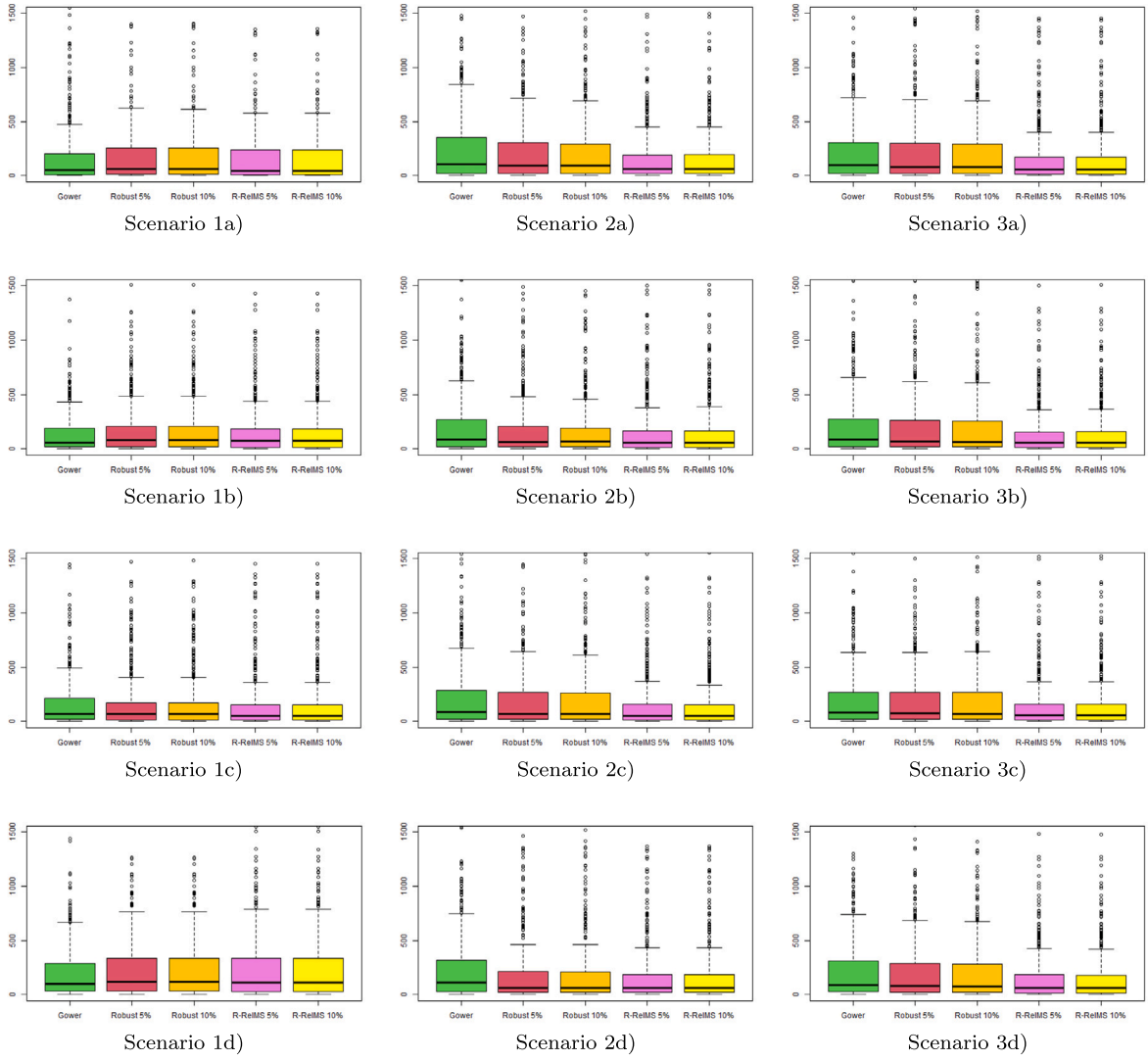

Fig. 6. SE distributions for DB-LM with Gower's, robust G-Gower (5%, 10% trimmed) and robust RelMS (5%, 10% trimmed) metrics, scenarios (1a) to (3d). Contaminated non-linear response.

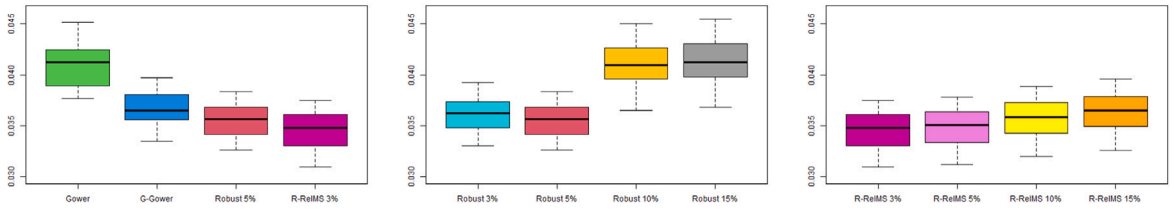

Fig. 7. SE distributions for DB-LM with Gower's, G-Gower, robust G-Gower (3%–15% trimmed) and robust RelMS (3%–15% trimmed) metrics. Bike sharing dataset.

Table 3
Estimated SE (mean, median and SD values) for the DB-LM with different metrics. For better comparison, values are standardized with respect to the minimum SE Mean value. Motor Insurance dataset.

Metric	SE - Mean values[a]			
	$k = 3$	$k = 4$	$k = 5$	$k = 6$
Euclidean	1.7653	2.9164	8.1313	8.1313
Gower	1.0045	1.0089	1.0049	1.0347
G-Gower	1.0001	1.0052	1.0096	1.0138
Robust G-Gower (3%)	1.0000	1.0051	1.0096	1.0137
Robust G-Gower (5%)	1.0001	1.0052	1.0097	1.0138
Robust G-Gower (10%)	1.0001	1.0052	1.0097	1.0138
Robust G-Gower (15%)	1.0001	1.0052	1.0097	1.0138
Robust RelMS (3%)	1.0004	1.0055	1.0099	1.0140
Robust RelMS (5%)	1.0004	1.0055	1.0099	1.0140
Robust RelMS (10%)	1.0004	1.0055	1.0099	1.0140
Robust RelMS (15%)	1.0004	1.0055	1.0099	1.0140
Metric	SE - Median values[a]			
	$k = 3$	$k = 4$	$k = 5$	$k = 6$
Euclidean	0.1744	0.1677	0.1677	0.1677
Gower	0.0839	0.0890	0.0933	0.1290
G-Gower	0.0721	0.0755	0.0756	0.0767
Robust G-Gower (3% trimmed)	0.0723	0.0755	0.0756	0.0766
Robust G-Gower (5% trimmed)	0.0721	0.0755	0.0755	0.0767
Robust G-Gower (10% trimmed)	0.0721	0.0755	0.0755	0.0766
Robust G-Gower (15% trimmed)	0.0722	0.0755	0.0755	0.0766
Robust RelMS (3% trimmed)	0.0705	0.0757	0.0792	0.0814
Robust RelMS (5% trimmed)	0.0708	0.0757	0.0792	0.0814
Robust RelMS (10% trimmed)	0.0708	0.0757	0.0791	0.0814
Robust RelMS (15% trimmed)	0.0707	0.0757	0.0791	0.0814
Metric	SE - SD values[a]			
	$k = 3$	$k = 4$	$k = 5$	$k = 6$
Euclidean	9.4523	14.3168	41.4739	41.4739
Gower	8.3846	8.4102	8.3638	8.3757
G-Gower	8.3923	8.3664	8.3876	8.4018
Robust G-Gower (3% trimmed)	8.3917	8.3662	8.3872	8.4015
Robust G-Gower (5% trimmed)	8.3922	8.3665	8.3876	8.4019
Robust G-Gower (10% trimmed)	8.3923	8.3665	8.3877	8.4019
Robust G-Gower (15% trimmed)	8.3922	8.3665	8.3876	8.4019
Robust RelMS (3% trimmed)	8.4030	8.3733	8.3946	8.4085
Robust RelMS (5% trimmed)	8.4021	8.3730	8.3943	8.4083
Robust RelMS (10% trimmed)	8.4021	8.3729	8.3943	8.4082
Robust RelMS (15% trimmed)	8.4022	8.3730	8.3944	8.4083

[a] Minimum SE Mean value 1085479.3501.

Table 4
Percentage of explained variability for $k = 3$ latent variables. Motor Insurance dataset.

Metric	`rel.gvar` for `eff.rank` = 3
Euclidean	99.99
Gower	64.92
G-Gower	84.88
Robust G-Gower (3%)	85.54
Robust G-Gower 5%)	85.06
Robust G-Gower (10%)	85.13
Robust G-Gower (15%)	85.23
Robust RelMS (3%)	87.95
Robust RelMS (5%)	87.51
Robust RelMS (10%)	87.61
Robust RelMS (15%)	87.70

Table 5
Selection of the number of latent variables based on GVC, AIC and BIC criteria. Motor Insurance dataset.

Metric	GCV			
	$k=3$	$k=4$	$k=5$	$k=6$
Robust G-Gower (3%)	1076704.7603	1083201.9426	1090646.9258	1092800.0770
Robust G-Gower (5%)	1076795.9577	1083279.4823	1090721.7813	1092850.7494
Robust RelMS (3%)	1077321.9380	1083661.8078	1091105.5195	1093128.0293
Robust RelMS (5%)	1077307.2209	1083677.1446	1091118.3490	1093133.2411
Metric	AIC			
	$k=3$	$k=4$	$k=5$	$k=6$
Robust G-Gower (3%)	5386.6712	5388.4219	5390.4116	5390.9554
Robust G-Gower (5%)	5386.6961	5388.4430	5390.4318	5390.9691
Robust RelMS (3%)	5441.8897	5443.5966	5445.5851	5446.0934
Robust RelMS (5%)	5386.8362	5388.5513	5390.5390	5391.0453
Metric	BIC			
	$k=3$	$k=4$	$k=5$	$k=6$
Robust G-Gower (3%)	5401.4191	5406.8568	5412.5334	5416.7642
Robust G-Gower (5%)	5401.4440	5406.8779	5412.5537	5416.7779
Robust RelMS (3%)	5456.6376	5462.0315	5467.7069	5471.9022
Robust RelMS (5%)	5401.5841	5406.9862	5412.6609	5416.8541

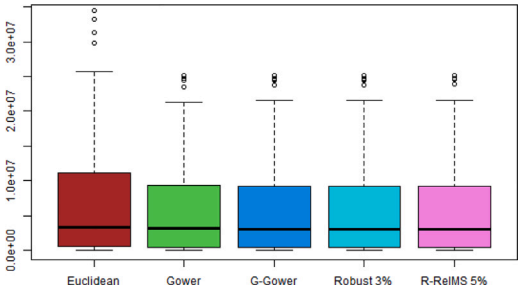

Fig. 8. SE distributions for DB-LM with Euclidean, Gower's, G-Gower, robust G-Gower (3% trimmed) and robust RelMS (5% trimmed) metrics. Motor Insurance dataset.

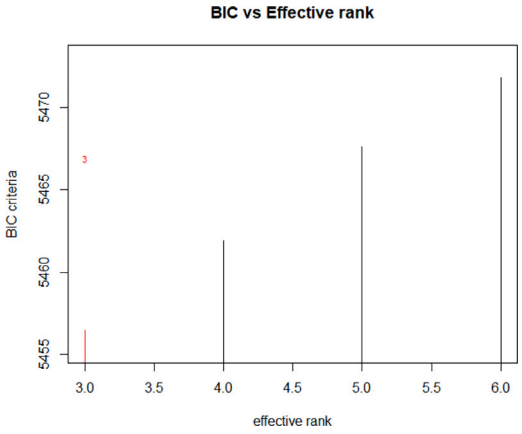

Fig. 9. Graphical output of `dbstats` for the optimal BIC for $k = 3, 4, 5, 6$, for the Robust DB model with G-Gower's (3% trimming). Motor Insurance dataset.

of DB-LM, and a very competitive model to be used, for instance, for classification purposes.

Funding

The authors acknowledge the support of grants PID2021-123592OB-I00, funded by MCIN/ AEI/10.13039/501100011033 and, as appropriate, by "ERDF A way of making Europe" and TED2021-129316B-I00, funded by MCIN/AEI/10.13039/501100011033 and, as appropriate, by the "European Union NextGenerationEU/PRTR".

CRediT authorship contribution statement

Eva Boj: Conceptualization, Methodology, Software, Data curation, Writing – original draft, Investigation. **Aurea Grané:** Conceptualization, Methodology, Data curation, Writing – original draft, Writing – review & editing, Investigation.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

Real datasets are available from the referred websites or R-packages.

Appendix A

In this section we include the usage of function `dblm` of `dbstats` package for R, used in the simulation study of Section 3.

```
## S3 method for class 'formula'
dblm(formula,data,...,metric="euclidean",method="OCV",
       full.search=TRUE, weights,rel.gvar=0.95,eff.rank)

## S3 method for class 'dist'
dblm(distance,y,...,method="OCV",full.search=TRUE,
       weights,rel.gvar=0.95,eff.rank)

## S3 method for class 'D2'
dblm(D2,y,...,method="OCV",full.search=TRUE,weights,rel.gvar=0.95,
       eff.rank)

## S3 method for class 'Gram'
dblm(G,y,...,method="OCV",full.search=TRUE,weights,rel.gvar=0.95,
       eff.rank)
```

Appendix B

In this section we include additional summary statistics concerning the simulation study presented in Section 3.

Appendix C

In this section a summary of the two best models fitted in Section 4 is provided.

Bike Sharing dataset

Table 6
Estimated SE (mean and median values) for the DB-LM with different metrics. Uncontaminated linear response.

SE - Mean values									
Scenario	Euclidean	Gower	G-Gower	Robust G-Gower			Robust RelMS		
				(5%)	(10%)	(15%)	(5%)	(10%)	(15%)
No outlier	1.9580	17.0233	2.8106	2.4996	2.5752	2.4862	1.3210	1.3235	1.3193
1 (a)	56.2766	19.5313	19.5953	19.4949	19.6114	19.6127	19.9681	19.9582	19.9561
1 (b)	1.9314	26.9541	1.1467	1.1445	1.1417	1.1414	1.3348	1.3326	1.3335
1 (c)	1.9649	17.0330	3.4566	3.0776	2.9977	3.0314	1.5631	1.5592	1.5606
1 (d)	55.4866	22.9153	17.6539	17.6596	17.6632	17.6640	17.7460	17.7382	17.7361
No outlier	2.2324	19.4220	6.8783	7.4542	6.2207	6.0889	1.2169	1.1971	1.1949
2 (a)	20.8702	22.2078	11.8075	13.1920	12.5830	12.2585	11.7435	11.7502	11.7525
2 (b)	2.1916	23.0635	2.1196	2.4859	1.9747	1.8214	1.2595	1.2496	1.2462
2 (c)	2.2219	19.5431	7.1502	8.0085	6.7605	6.2859	1.4178	1.3730	1.3591
2 (d)	20.8555	27.7483	10.2229	10.2926	10.2605	10.2476	10.9419	10.9483	10.9483
No outlier	2.2324	19.4220	6.8783	7.4542	6.2207	6.0889	1.2169	1.1971	1.1949
3 (a)	4.6395	19.2883	6.8663	10.3831	9.9687	9.3934	3.0310	3.0283	3.0256
3 (b)	2.2033	19.3972	6.6076	7.3602	6.2401	5.8301	1.2017	1.1893	1.1853
3 (c)	2.2125	18.9615	7.4799	8.3638	7.0419	6.5129	1.3012	1.2728	1.2638
3 (d)	4.5856	19.6338	5.8029	8.4498	8.0959	7.6241	2.8186	2.8024	2.7814
SE - Median values									
Scenario	Euclidean	Gower	G-Gower	Robust G-Gower			Robust RelMS		
				(5%)	(10%)	(15%)	(5%)	(10%)	(15%)
No outlier	1.0952	6.3648	1.0570	0.9947	0.9874	1.0109	0.5041	0.5057	0.5058
1 (a)	24.0337	7.4969	8.1838	8.2038	8.1966	8.1904	8.0648	8.0723	8.0736
1 (b)	1.0655	8.7608	0.4872	0.4950	0.4930	0.4961	0.5756	0.5775	0.5794
1 (c)	1.0953	6.5447	1.5327	1.3649	1.3387	1.3531	0.6195	0.6159	0.6161
1 (d)	23.6448	5.5489	5.1411	5.0996	5.1008	5.0928	4.9945	4.9822	4.9893
No outlier	1.0097	8.1789	3.4087	3.7769	3.1287	3.0355	0.5359	0.5097	0.5127
2 (a)	4.4264	9.5153	3.4076	5.2418	4.4880	4.0564	3.0013	2.9892	3.0090
2 (b)	0.9777	9.1180	0.8979	1.0009	0.8834	0.8434	0.5598	0.5608	0.5573
2 (c)	0.9936	8.1005	3.6028	4.1721	3.3273	3.1425	0.6283	0.6080	0.5971
2 (d)	4.4749	10.7503	2.5945	2.7363	2.5790	2.5496	2.6450	2.6243	2.6427
No outlier	1.0097	8.1789	3.4087	3.7769	3.1287	3.0355	0.5359	0.5097	0.5127
3 (a)	1.7730	8.2206	3.1175	5.1673	5.0041	4.6478	1.1434	1.1302	1.1413
3 (b)	0.9868	7.8193	3.4245	3.8763	3.2663	2.9949	0.5429	0.5512	0.5499
3 (c)	0.9950	7.9684	3.8958	4.1752	3.7131	3.4262	0.5526	0.5553	0.5432
3 (d)	1.7485	8.0157	2.6350	3.6923	3.5124	3.3808	1.1194	1.1107	1.1122

Table 7
Estimated SE (mean and median values) for the DB-LM with different metrics. Contaminated linear response.

SE - Mean values									
Scenario	Euclidean	Gower	G-Gower	Robust G-Gower			Robust RelMS		
				(5%)	(10%)	(15%)	(5%)	(10%)	(15%)
No outlier	1.9580	17.0233	2.8106	2.4996	2.5752	2.4862	1.3210	1.3235	1.3193
1 (a)	2.8595	16.3084	1.0092	0.9911	0.9737	0.9718	1.1793	1.1619	1.1586
1 (b)	2.2187	24.3870	0.9833	0.9811	0.9794	0.9793	1.1365	1.1346	1.1351
1 (c)	2.0919	16.5985	3.4304	3.0538	2.9749	3.0090	1.2539	1.2510	1.2521
1 (d)	3.9314	7.1985	1.0908	1.0477	1.0367	1.0301	1.0682	1.0573	1.0551
No outlier	2.2324	19.4220	6.8783	7.4542	6.2207	6.0889	1.2169	1.1971	1.1949
2 (a)	2.5550	47.2122	2.7358	12.1241	8.5426	6.2762	1.2907	1.2672	1.2486
2 (b)	2.2631	21.9759	1.8717	2.2021	1.7442	1.6087	1.1371	1.1279	1.1250
2 (c)	2.2478	19.1187	6.9756	7.8114	6.5956	6.1335	1.2355	1.1906	1.1764
2 (d)	2.8088	60.5340	1.2730	2.3900	1.9740	1.7451	1.1576	1.1372	1.1238
No outlier	2.2324	19.4220	6.8783	7.4542	6.2207	6.0889	1.2169	1.1971	1.1949
3 (a)	2.1988	22.7100	6.3502	10.5280	10.0397	9.2522	1.1953	1.1904	1.1831
3 (b)	2.2105	19.2463	6.5413	7.2875	6.1766	5.7711	1.1681	1.1562	1.1526
3 (c)	2.2111	18.9234	7.4802	8.3630	7.0433	6.5160	1.2062	1.1825	1.1750
3 (d)	2.2705	23.9345	5.1215	8.3753	7.9129	7.2604	1.2282	1.2148	1.1968
SE - Median values									
Scenario	Euclidean	Gower	G-Gower	Robust G-Gower			Robust RelMS		
				(5%)	(10%)	(15%)	(5%)	(10%)	(15%)
No outlier	1.0952	6.3648	1.0570	0.9947	0.9874	1.0109	0.5041	0.5057	0.5058
1 (a)	1.4632	2.7432	0.4687	0.4592	0.4431	0.4448	0.4698	0.4638	0.4561
1 (b)	1.1449	8.4370	0.4127	0.4181	0.4100	0.4096	0.4737	0.4721	0.4720
1 (c)	1.1072	6.3285	1.3101	1.2001	1.1710	1.1868	0.4633	0.4637	0.4607
1 (d)	2.2851	1.2764	0.4747	0.4514	0.4482	0.4457	0.4129	0.4058	0.4115

(*continued on next page*)

Table 7 (*continued*).

No outlier	1.0097	8.1789	3.4087	3.7769	3.1287	3.0355	0.5359	0.5097	0.5127
2 (a)	1.0699	12.0355	1.2796	5.0200	3.5157	2.5504	0.5557	0.5403	0.5305
2 (b)	1.0198	8.9597	0.8576	0.9652	0.8620	0.7788	0.5317	0.5254	0.5261
2 (c)	0.9690	7.6475	3.8032	4.1321	3.4794	3.1549	0.5660	0.5364	0.5322
2 (d)	1.0709	12.0573	0.6132	1.0400	0.9223	0.8370	0.5457	0.5378	0.5202
No outlier	1.0097	8.1789	3.4087	3.7769	3.1287	3.0355	0.5359	0.5097	0.5127
3 (a)	1.0344	9.2760	3.1695	5.1970	4.9968	4.3832	0.5342	0.5342	0.5383
3 (b)	1.0091	7.9243	3.5051	3.9283	3.1627	2.9722	0.5399	0.5385	0.5350
3 (c)	1.0006	7.9867	3.8574	4.0914	3.7382	3.4111	0.5368	0.5613	0.5533
3 (d)	1.0724	9.7907	2.7814	4.3576	4.0048	3.6427	0.5249	0.5393	0.5269

Table 8
Estimated SE (mean and median values) for the DB-LM with different metrics. Uncontaminated non-linear response.

SE - Mean values

Scenario	Euclidean	Gower	G-Gower	Robust G-Gower			Robust RelMS		
				(5%)	(10%)	(15%)	(5%)	(10%)	(15%)
No outlier	238.3173	157.1809	155.1188	154.0918	154.3570	154.0608	133.8524	133.8154	133.8872
1 (a)	299.8464	159.5251	173.1662	172.7206	173.1613	173.1585	170.2995	170.3866	170.3971
1 (b)	237.3054	155.9672	145.4180	145.3795	145.3662	145.3682	136.1690	136.1978	136.1851
1 (c)	237.2702	181.3580	164.4569	162.9331	162.6195	162.7557	146.0514	146.0716	146.0704
1 (d)	296.3721	169.8422	184.9105	184.8163	184.7964	184.7758	178.9406	179.0016	179.0142
No outlier	218.4939	226.1340	199.2341	201.7821	196.6168	196.0155	155.5568	156.1126	156.1677
2 (a)	236.1479	236.1273	192.9356	204.7988	200.6031	197.9733	184.6481	184.8545	185.0808
2 (b)	214.0613	225.7478	169.9376	171.9729	169.5626	168.8704	157.5916	157.9183	158.0297
2 (c)	213.7543	232.9335	200.5033	204.6432	199.0513	196.5159	154.7166	155.5569	155.8759
2 (d)	235.9091	244.6872	191.7630	193.1335	192.8574	192.7281	189.8525	189.8564	189.9269
No outlier	218.4939	226.1340	199.2341	201.7821	196.6168	196.0155	155.5568	156.1126	156.1677
3 (a)	217.8814	224.2229	194.4221	209.7902	208.6269	206.5369	159.6568	159.7351	159.9254
3 (b)	213.6699	223.9224	196.0288	199.0313	194.8861	193.0508	155.1689	155.5982	155.7480
3 (c)	213.6684	224.0319	198.7451	202.5439	197.2227	194.6475	153.6600	154.2439	154.4631
3 (d)	218.2937	224.7703	191.2000	202.7763	201.5916	199.8082	165.9616	165.8027	165.5866

SE - Median values

Scenario	Euclidean	Gower	G-Gower	Robust G-Gower			Robust RelMS		
				(5%)	(10%)	(15%)	(5%)	(10%)	(15%)
No outlier	123.3531	50.0488	67.5421	66.1134	66.5338	66.0257	51.8302	51.9325	51.8763
1 (a)	154.8938	54.8426	66.6913	67.1975	66.7830	66.7634	63.6243	63.9068	64.0383
1 (b)	121.7133	51.6823	62.2546	62.2903	62.2253	62.2890	56.1822	56.1660	56.2701
1 (c)	121.5676	60.0283	70.6065	70.6409	69.4756	70.2544	60.2813	60.0902	59.9119
1 (d)	154.1955	55.6864	76.0208	75.2593	74.8426	74.7895	73.8104	74.0328	74.0538
No outlier	88.3718	86.0277	65.5242	68.1828	62.3379	61.6429	53.6562	52.8028	52.7245
2 (a)	93.9194	86.0255	65.5756	77.7447	74.7492	73.5954	65.8480	65.6486	65.6767
2 (b)	88.8261	84.3103	65.8192	64.4552	66.1555	66.3042	55.9639	55.8997	56.0139
2 (c)	88.2620	85.2808	66.0798	67.1100	64.9764	65.1332	55.6536	56.4550	56.5542
2 (d)	93.2571	87.8864	68.2743	65.3360	66.9787	66.4696	64.7444	64.4468	64.0736
No outlier	88.3718	86.0277	65.5242	68.1828	62.3379	61.6429	53.6562	52.8028	52.7245
3 (a)	87.8727	83.1122	67.9923	79.1322	76.1739	75.3826	55.1581	55.1785	55.2259
3 (b)	88.1882	86.5779	65.9046	70.1526	64.3658	63.6876	55.9725	55.8431	55.8265
3 (c)	88.2662	81.3862	69.9464	74.3170	67.0350	63.5721	54.2485	54.5572	54.8646
3 (d)	87.5141	83.7526	69.5614	71.5030	71.3192	70.1975	59.5665	59.8736	60.1798

```
R> dblmggower_rob05 <- dblm(Dggower_rob05, Response, method = "rel.gvar",
 rel.gvar = 0.5)

R> summary(dblmggower_rob05)

call:   dblm(D2 = Dggower_rob05, y = Response, method = "rel.gvar", rel.gvar = 0.5)

Weighted Residuals:
    Min.  1st Qu.   Median     Mean  3rd Qu.     Max.
-1.14500 -0.11501  0.00942  0.00000  0.12272  0.84969

R-squared: 0.754931   Adjusted R-squared: 0.754169
Weighted Geometric Variability: 1.000000

Used effective rank = 9
Relative geometric variability = 0.525975

R> dblmggower_rob_Relmds03 <- dblm(Dggower_rob_Relmds03, Response,
 method = "rel.gvar", rel.gvar = 0.5)

R> summary(dblmggower_rob_Relmds03)

call:   dblm(D2 = Dggower_rob_Relmds03, y = Response, method = "rel.gvar",
    rel.gvar = 0.5)

Weighted Residuals:
     Min.   1st Qu.    Median      Mean   3rd Qu.      Max.
-0.868893 -0.071424  0.000574  0.000000  0.079289  0.611387

R-squared: 0.883529   Adjusted R-squared: 0.786077
Weighted Geometric Variability: 0.004266

Used effective rank = 1322
Relative geometric variability = 0.500106
```

Motor Insurance dataset

```
R> dblmggower <- dblm(Dggower, y, weights = w,  method = "eff.rank", eff.rank = 3)
R> summary(dblmggower)

call:   dblm(D2 = Dggower, y = y, method = "eff.rank", weights = w, eff.rank = 3)

Weighted Residuals:
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
 -18728   -6054   -1284     456    5613   37568

R-squared: 0.048320   Adjusted R-squared: 0.038508
Weighted Geometric Variability: 0.651663

Used effective rank = 3
Relative geometric variability = 0.848801

R> dblmggower_rob03 <- dblm(Dggower_rob03, y, weights = w, method = "eff.rank",
 eff.rank = 3)
R> summary(dblmggower_rob03)

call:   dblm(D2 = Dggower_rob03, y = y, method = "eff.rank", weights = w,
    eff.rank = 3)
```

Table 9
Estimated SE (mean and median values) for the DB-LM with different metrics. Contaminated non-linear response.

SE - Mean values									
Scenario	Euclidean	Gower	G-Gower	Robust G-Gower			Robust RelMS		
				(5%)	(10%)	(15%)	(5%)	(10%)	(15%)
No outlier	238.3173	157.1809	155.1188	154.0918	154.3570	154.0608	133.8524	133.8154	133.8872
1 (a)	554.5634	282.0960	444.3794	447.7025	443.5773	443.5119	407.8975	409.3240	409.5465
1 (b)	238.6875	153.7460	165.1816	165.1608	165.1617	165.1659	157.0000	157.0192	157.0081
1 (c)	235.8697	174.6556	168.4490	167.2659	167.0327	167.1438	146.9566	147.0011	146.9834
1 (d)	652.4852	501.0114	534.3033	534.3620	534.3812	534.3907	533.7185	533.8767	533.8947
No outlier	218.4939	226.1340	199.2341	201.7821	196.6168	196.0155	155.5568	156.1126	156.1677
2 (a)	278.8127	394.8997	235.9094	281.7267	265.5537	254.9256	212.1332	212.4572	212.8673
2 (b)	215.6545	221.0478	174.0044	175.5232	173.8079	173.3347	164.0965	164.4029	164.5091
2 (c)	213.6516	232.9090	201.5071	205.5283	200.0497	197.5783	154.7521	155.5859	155.9039
2 (d)	317.7182	426.5446	303.0997	303.2137	303.6806	304.0067	290.1317	290.5296	290.9680
No outlier	218.4939	226.1340	199.2341	201.7821	196.6168	196.0155	155.5568	156.1126	156.1677
3 (a)	225.7202	256.6361	214.3299	234.0348	232.0263	228.7474	167.7205	167.8369	168.0924
3 (b)	213.7139	223.8310	195.7568	198.7732	194.6097	192.7642	154.8060	155.2351	155.3844
3 (c)	214.7464	225.3363	201.0275	204.7423	199.5131	196.9674	154.5541	155.1755	155.4079
3 (d)	224.8489	255.9227	208.4456	222.8559	221.1599	218.4654	174.1653	174.0238	173.8398
SE - Median values									
Scenario	Euclidean	Gower	G-Gower	Robust G-Gower			Robust RelMS		
				(5%)	(10%)	(15%)	(5%)	(10%)	(15%)
No outlier	123.3531	50.0488	67.5421	66.1134	66.5338	66.0257	51.8302	51.9325	51.8763
1 (a)	132.2787	49.2286	60.2295	59.8209	59.0006	58.8267	44.8619	45.1427	44.8210
1 (b)	131.1358	58.0541	78.7007	78.7523	78.6505	78.6597	75.1920	75.2184	75.2201
1 (c)	117.3067	64.2930	69.1175	68.8134	68.8957	68.6412	50.6489	50.6430	50.6507
1 (d)	168.3795	97.7283	112.8336	113.2379	113.2043	113.3078	109.0198	109.0048	109.1882
No outlier	88.3718	86.0277	65.5242	68.1828	62.3379	61.6429	53.6562	52.8028	52.7245
2 (a)	99.2445	103.9791	77.4780	89.9221	88.9706	85.5208	63.0179	63.3059	63.5843
2 (b)	84.4925	85.8134	64.0143	62.3142	63.9714	64.0712	57.4281	57.4219	57.4539
2 (c)	83.9401	85.7369	67.6902	67.9126	66.5492	65.5840	50.3446	51.2212	51.6233
2 (d)	78.8361	109.6602	62.6724	62.8228	63.2006	63.7499	61.5119	61.2971	60.9274
No outlier	88.3718	86.0277	65.5242	68.1828	62.3379	61.6429	53.6562	52.8028	52.7245
3 (a)	88.3744	95.0482	74.6332	77.1671	79.7410	76.3644	54.8303	54.9475	55.1147
3 (b)	87.0042	85.6743	65.4463	69.2360	63.9322	63.6203	55.8833	55.8584	55.8226
3 (c)	85.3812	81.9391	67.3783	72.1421	67.0824	65.0114	53.3714	54.3489	53.9441
3 (d)	87.8703	84.8822	68.5795	78.3703	75.5568	73.7965	58.6059	58.5713	57.8366

Table 10
Estimated SE SD values for the DB-LM with different metrics. Uncontaminated linear response.

SE - SD values									
Scenario	Euclidean	Gower	G-Gower	Robust G-Gower			Robust RelMS		
				(5%)	(10%)	(15%)	(5%)	(10%)	(15%)
No outlier	2.4369	36.5427	4.2843	3.7939	3.9170	3.7767	2.1003	2.1033	2.0928
1 (a)	99.5591	37.0878	33.9527	33.8489	33.9736	33.9754	34.6308	34.6184	34.6119
1 (b)	2.3899	64.3067	1.5891	1.5856	1.5813	1.5806	1.9906	1.9867	1.9884
1 (c)	2.4224	36.2553	5.2445	4.6585	4.5346	4.5916	2.5261	2.5168	2.5193
1 (d)	98.9138	60.0598	44.2665	44.3701	44.4046	44.4234	44.7141	44.6814	44.6754
No outlier	3.0476	28.4076	9.2582	10.0500	8.4026	8.2252	1.6971	1.6757	1.6730
2 (a)	80.9630	32.1341	34.8541	24.1602	26.9158	29.4690	36.9050	37.1085	37.2506
2 (b)	3.0131	41.3029	3.6193	4.3586	3.3215	3.0124	1.8018	1.7870	1.7813
2 (c)	3.0376	28.0605	9.2456	10.3886	8.7578	8.1435	2.2136	2.1781	2.1686
2 (d)	80.6868	58.3941	31.4632	29.6451	30.2404	30.6150	34.4525	34.5874	34.6747
No outlier	3.0476	28.4076	9.2582	10.0500	8.4026	8.2252	1.6971	1.6757	1.6730
3 (a)	7.8182	27.9715	8.9961	13.7812	13.2021	12.3823	5.8752	5.8752	5.8765
3 (b)	3.0177	27.8335	8.5953	9.5902	8.1336	7.6039	1.6586	1.6425	1.6368
3 (c)	3.0262	27.5359	9.8979	11.1436	9.3180	8.5947	1.8136	1.7781	1.7663
3 (d)	7.7316	28.4051	7.9016	11.1807	10.7236	10.1281	5.5355	5.5185	5.4978

Table 11
Estimated SE SD values for the DB-LM with different metrics. Contaminated linear response.

SE - SD values									
Scenario	Euclidean	Gower	G-Gower	Robust G-Gower			Robust RelMS		
				(5%)	(10%)	(15%)	(5%)	(10%)	(15%)
No outlier	2.4369	36.5427	4.2843	3.7939	3.9170	3.7767	2.1003	2.1033	2.0928
1 (a)	4.1238	37.5187	1.4480	1.4041	1.3902	1.3873	1.8045	1.7675	1.7572
1 (b)	3.2511	55.7483	1.3659	1.3632	1.3618	1.3618	1.7096	1.7054	1.7055
1 (c)	2.8420	35.5106	5.2199	4.6270	4.5011	4.5587	2.0031	1.9969	1.9986
1 (d)	5.2991	23.7958	1.5461	1.4766	1.4602	1.4499	1.5758	1.5565	1.5520
No outlier	3.0476	28.4076	9.2582	10.0500	8.4026	8.2252	1.6971	1.6757	1.6730
2 (a)	3.7830	154.8224	3.7561	21.9191	13.8208	9.3906	2.0715	1.9934	1.9335
2 (b)	3.0906	35.9296	2.6069	3.0963	2.4199	2.2226	1.5934	1.5807	1.5762
2 (c)	3.0828	27.7300	9.1943	10.3392	8.7047	8.0830	1.7015	1.6483	1.6320
2 (d)	4.9504	253.1834	1.7280	3.5337	2.7596	2.3947	1.7797	1.7104	1.6658
No outlier	3.0476	28.4076	9.2582	10.0500	8.4026	8.2252	1.6971	1.6757	1.6730
3 (a)	3.0305	33.3477	8.3783	14.1247	13.4516	12.3974	1.7135	1.7050	1.6918
3 (b)	3.0176	27.5573	8.5967	9.5835	8.1371	7.6075	1.6286	1.6132	1.6080
3 (c)	3.0283	27.5472	9.9634	11.2170	9.3771	8.6471	1.6628	1.6335	1.6244
3 (d)	3.1057	36.9521	6.6602	10.8495	10.2536	9.4344	1.7573	1.7359	1.7058

Table 12
Estimated SE SD values for the DB-LM with different metrics. Uncontaminated non-linear response.

SE - SD values									
Scenario	Euclidean	Gower	G-Gower	Robust G-Gower			Robust RelMS		
				(5%)	(10%)	(15%)	(5%)	(10%)	(15%)
No outlier	410.4110	341.5664	276.6520	274.7538	275.3178	274.8358	252.1054	252.0375	252.1751
1 (a)	480.3126	323.0275	321.6728	321.3294	321.6119	321.6143	316.1344	316.2425	316.2459
1 (b)	407.0011	323.1971	257.3756	257.3898	257.4316	257.4568	245.4684	245.5230	245.5060
1 (c)	406.9019	371.3920	276.2363	274.4998	274.3225	274.5533	268.5309	268.6170	268.6138
1 (d)	474.6929	345.3905	334.9428	334.9237	334.8848	334.8875	323.3328	323.3959	323.4167
No outlier	397.6678	364.2506	308.8336	311.9782	306.4142	305.7433	286.1494	286.7466	286.7848
2 (a)	366.3723	379.2846	297.7805	311.3156	304.6721	301.1842	293.7282	294.0861	294.4162
2 (b)	395.3036	358.6080	289.6459	289.0874	290.1187	290.5202	286.5175	287.0004	287.2240
2 (c)	395.3204	375.2875	314.0983	319.0799	312.9316	310.3186	291.8898	292.8470	293.2520
2 (d)	366.3277	399.4877	304.5377	301.9436	302.8515	303.5500	303.4681	303.6224	303.7939
No outlier	397.6678	364.2506	308.8336	311.9782	306.4142	305.7433	286.1494	286.7466	286.7848
3 (a)	384.9863	358.8436	301.6428	322.1438	320.3871	317.5380	278.7906	278.9074	279.2171
3 (b)	395.3324	355.8383	303.9895	307.2463	303.2405	301.5963	285.3593	285.9061	286.1569
3 (c)	395.2893	356.3250	305.7642	310.5427	304.4353	301.7326	284.6054	285.2881	285.5961
3 (d)	385.0498	355.9146	296.3709	309.7553	308.3047	306.4236	280.0942	280.1193	280.2584

Table 13
Estimated SE SD values for the DB-LM with different metrics. Non-contaminated non-linear response.

SE - SD values									
Scenario	Euclidean	Gower	G-Gower	Robust G-Gower			Robust RelMS		
				(5%)	(10%)	(15%)	(5%)	(10%)	(15%)
No outlier	410.4110	341.5664	276.6520	274.7538	275.3178	274.8358	252.1054	252.0375	252.1751
1 (a)	1389.3886	753.7295	1331.5333	1357.4035	1335.0443	1335.2742	1262.0137	1265.7996	1266.4750
1 (b)	389.6733	320.9154	294.8955	294.7933	294.8127	294.8463	275.4810	275.5078	275.4867
1 (c)	404.6081	365.9163	296.0253	293.9168	293.6731	293.9518	278.9747	279.0431	279.0021
1 (d)	1875.0122	1548.6199	1598.3354	1598.4075	1598.4196	1598.4388	1597.3180	1597.6821	1597.7075
No outlier	397.6678	364.2506	308.8336	311.9782	306.4142	305.7433	286.1494	286.7466	286.7848
2 (a)	696.7287	1531.7533	650.1827	842.9077	765.0771	715.7555	593.3202	593.8031	594.6338
2 (b)	392.0892	357.3558	295.4163	294.4111	296.1005	296.7432	297.0986	297.4944	297.6887
2 (c)	398.4928	376.6503	315.6249	320.5477	314.4219	311.9001	295.1683	296.1724	296.5916
2 (d)	1135.0675	1723.3369	1399.8033	1307.2651	1338.2538	1356.8192	1333.8906	1340.7895	1345.2792
No outlier	397.6678	364.2506	308.8336	311.9782	306.4142	305.7433	286.1494	286.7466	286.7848
3 (a)	424.2297	453.5577	353.3837	381.8992	378.2835	372.4940	320.2327	320.4497	320.9285
3 (b)	394.9026	355.9919	303.6696	306.9586	302.9063	301.2460	284.9851	285.5320	285.7833
3 (c)	399.3096	369.9716	316.9492	322.1324	315.5196	312.5349	290.5824	291.3097	291.6294
3 (d)	450.5049	532.1365	405.3148	427.5591	424.6896	419.8443	358.4244	358.3291	358.2881

```
Weighted Residuals:
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
 -18724   -6051   -1285     455    5610   37567

R-squared: 0.048381   Adjusted R-squared: 0.038570
Weighted Geometric Variability: 0.681554

Used effective rank = 3
Relative geometric variability = 0.855434
```

References

[1] Borg I, Groenen P. Modern multidimensional scaling: theory and applications. New York: Springer; 1997.

[2] Cuadras CM, Dodge Y. Distance analysis in discrimination and classification using both continuous and categorical variables. In: Statistical data analysis and inference. Amsterdam: North-Holland; 1989, p. 459–73.

[3] Cuadras CM, Arenas C. A distance-based model for prediction with mixed data. Commun Stat - Theory Methods 1990;19:2261–79.
[4] Cuadras CM, Arenas C, Fortiana J. Some computational aspects of a distance-based model for prediction. Comm Statist Simulation Comput 1996;25(3):593–609.
[5] Boj E, Claramunt MM, Fortiana J. Selection of predictors in distance-based regression. Commun Stat - Theory Methods 2007;36:87–98.
[6] Esteve A, Boj E, Fortiana J. Interaction terms in distance-based regression. Comm Statist Theory Methods 2009;38(3498):19–3509.
[7] Boj E, Delicado P, Fortiana J. Local linear functional regression based on weighted distance-based regression. Comput Stat Data Anal 2010;54:429–37.
[8] Boj E, Delicado P, Fortiana J, Esteve A, Caballé A. Global and local distance-based generalized linear models. TEST 2016;25:170–95.
[9] Boj E, Caballé A, Delicado P, Fortiana J. Dbstats: distance-based statistics (dbstats). r package version 2.0.2. 2024, http://CRAN.R-project.org/package=dbstats.
[10] Baíllo A, Grané A. Subsampling and aggregation: A solution to the scalability problem in distance-based prediction for mixed-type data. Mathematics 2021;9:2247.
[11] Gower JC. A general coefficient of similarity and some of its properties. Biometrika 1971;27:857–74.
[12] Grané A, Salini S, Verdolini E. Robust multivariate analysis for mixed-type data: Novel algorithm and its practical application in socio-economic research. Socio-Econ Plan Sci 2021;73:100907.
[13] R Development Core Team. R: A Language and Environment for Statistical Computing. Vienna, Austria; 2024, http://www.R-project.org/.
[14] Gower JC, Hand DJ. Biplots. London: Chapman and Hall; 1996.
[15] Grané A, Manzi G, Salini S. Smart visualization of mixed data. Stats 2021;4:472–85.
[16] Cuadras CM, Fortiana J. Visualizing categorical data with related metric scaling. In: Blasius J, Greenacre M, editors. Visualization of Categorical Data. London: Academic Press; 1998.
[17] Albarrán I, Alonso P, Grané A. Profile identification via weighted related metric scaling: an application to dependent spanish children. J. R. Stat. Soc. - Ser. A. Stat. Soc. 2015;178:1–26.
[18] Grané A, Romera R. On visualizing mixed-type data: A joint metric approach to profile construction and outlier detection. Sociol Methods Res 2018;47(2):207–39.
[19] Cerioli A, Riani M, Atkinson AC, Corbellini A. The power of monitoring: how to make the most of a contaminated multivariate sample. Stat Methods Appl 2018;27:559–87.
[20] Hallin M, Ingenbleek JF. The Swedish automobile portfolio in 1977. a statistical study. Skandinavisk Aktuarietidskrift (Scand. Actuarial J.) 1983;83:49–64.
[21] Faraway J. Faraway: functions and datasets for books by Julian Faraway. 2022, R package version 1.0.8, http://CRAN.R-project.org/package=faraway.

Eva Boj is a Professor at the Department of Economic, Financial and Actuarial Mathematics of the University of Barcelona (Spain). She received Ph.D. in Actuarial and Financial Sciences at University of Barcelona (2003). Her fields of interest are multivariate analysis, actuarial sciences and in general data science. Since 2006 she has participated in several national research projects devoted to non-parametric techniques based on distances with application to complex data. She has been the Coordinator of the Working Group on Multivariate Analysis and Classification of the Spanish Society of Statistics from 2017 to 2022, in the International Federation Classification Societies.

Aurea Grané is Full Professor of Statistics at Universidad Carlos III de Madrid (Spain). She received Ph.D. in Mathematics at University of Barcelona (1999). Her fields of interest are multivariate and functional data analysis, outlier detection, goodness-of-fit tests and data science. Since 2010 she has led national research projects whose objectives were the development of non-parametric techniques based on distances with application to data of a certain complexity, and collaborated in European research projects. She is the current coordinator of the Working Group on Multivariate Analysis and Classification of the Spanish Society of Statistics.