INFORMATICA, 2016, Vol. 27, No. 4, 893–910
© 2016 *Vilnius University*
DOI: http://dx.doi.org/10.15388/Informatica.2016.116

A Hybrid Computational Method Based on Convex Optimization for Outlier Problems: Application to Earthquake Ground Motion Prediction

Fatma YERLİKAYA-ÖZKURT[1]*, Aysegul ASKAN[2],
Gerhard-Wilhelm WEBER[1]

[1] *Institute of Applied Mathematics, Middle East Technical University, 06800 Ankara, Turkey*
[2] *Department of Civil Engineering, Middle East Technical University, 06800 Ankara, Turkey*
e-mail: fatmayerlikaya@gmail.com, aaskan@metu.edu.tr, gweber@metu.edu.tr

Received: May 2015; accepted: March 2016

Abstract. Statistical modelling plays a central role for any prediction problem of interest. However, predictive models may give misleading results when the data contain outliers. In many real-world applications, it is important to identify and treat the outliers without direct elimination. To handle such issues, a hybrid computational method based on conic quadratic programming is introduced and employed on earthquake ground motion dataset. This method aims to minimize the impact of the outliers on regression estimators as well as handling the nonlinearity in the dataset. Results are compared against widely used parametric and nonparametric ground motion prediction models.

Key words: outlier detection procedure, mean shift outlier model, conic multivariate adaptive regression splines, ground motion prediction equations.

1. Introduction

Regression methods are well-known mathematical tools for investigating the relationship between a dependent variable and independent variable(s) (Montgomery and Peck, 1992). Among alternative regression methods, Linear Regression (LR) is usually preferred in many studies because of its well-established form and available computer packages. This method is based on certain assumptions which must be satisfied for valid results. However, in real-world applications, these assumptions are not always validated due to outliers in the data. Outliers, often seen as contamination to the data, reduce and affect the information that we may get from the source. Therefore, for any prediction problem of interest, it is essential to identify the existing outlier observations in the datasets (Barnett and Lewis, 1994).

During the construction of regression models on datasets with outliers, data transformation techniques are required. Such techniques may take time and typically need expertise. To handle these problems, robust statistical techniques which are not easily affected

*Corresponding author.

by outliers have been introduced (Rousseeuw and Leroy, 1987). Robust regression methods aim to minimize the impact of the outliers on regression estimators, but still invoke parametric assumptions after smoothing the influence of outliers on the regression line (Lane, 2002).

In most of the real-world applications in fields such as finance, medicine, and engineering, datasets contain outliers. Among these, one important specific application of modelling in the existence of outlier observations is earthquake ground motion prediction problem due to its random nature.

Earthquakes are among natural disasters with significant damage potential to urban areas all over the world due to the ground shakings involved. It is important to estimate potential ground motions due to possible future earthquakes for both seismic design and analysis purposes. Ground Motion Prediction Equations (GMPEs) are equations that employ empirical data and express certain peak ground motion parameters (e.g. Peak Ground Acceleration (PGA), Peak Ground Velocity (PGV) or spectral quantities) as functions of earthquake magnitude, source to site distance, site conditions at the stations and other physical parameters whenever available. Even though there is no physical form relating peak ground motion parameters to independent variables such as magnitude, distance and site conditions; most of the existing GMPEs are based on parametric regression techniques where the form of the predicting model is assigned a priori (e.g. Spudich *et al.*, 1999; Boore and Atkinson, 2008). Recently, non-parametric approaches for ground motion estimation have also been implemented (e.g. Alavi and Gandomi, 2011; Peruš and Fajfar, 2010; Tezcan and Cheng, 2012). In a previous study (Yerlikaya-Özkurt *et al.*, 2014), the authors have introduced a novel non-parametric ground motion prediction model with the use of Conic Multivariate Adaptive Regression Splines (CMARS).

Due to the inherent randomness in the temporal occurrences of earthquakes, the statistical models for inclusion of uncertainties are well suited for ground motion predictions. One of the most effective approaches is developed by Calvin and Žilinskas (2005). In this study, a statistical model of an objective function is minimized over a continuous interval to find the global minimum in the presence of uncertainties (outliers).

When the given problem originates from optimization, especially, from global optimization, then smart approaches from statistics (or approximation theory) and from optimization can be followed. In the paper by Žilinskas (2010), the author provides a careful introduction and important comparison of these approaches and introduces a new method based on data smoothing with radial basis function model and the least-squares approximation for noisy datasets. Based on this investigation, the author discloses a coincidence between the two models, that we are led to identical algorithms which combine advantages of statistical model and global optimization to find the best approximation in the presence of outlier observations.

As a result of increasing seismic station networks all over the world, recently the amount of ground motion data has increased. However, since catastrophic earthquakes are rare events in nature, ground motion data from moderate to large events are still rare and thus invaluable. This observation leads to ground motion models derived using all of the available data. Thus, for robust estimations, outlier analyses are required on ground motion datasets that do not lead to any loss of data. For this purpose, in this study we propose a computational approach that combines the robust outlier algorithm Mean Shift Outlier Model (MSOM) (Cook, 1982; Kim *et al.*, 2008) with CMARS (Weber *et al.*, 2012) to handle the nonlinearity in the dataset.

Our study and similar studies (Calvin and Žilinskas, 2005; Žilinskas, 2010) based on statistical models and global optimization algorithms are crucial, since the ground motion prediction problem remains at the heart of multiple disciplines ranging from earthquake engineering to risk management.

The structure of this article is organized as follows: in Section 2, the proposed methodology is presented. Section 3 presents the ground motion dataset used in this study followed by Section 4 which describes the specific application of the presented method. The same section also includes comparison of our results against existing alternatives. Finally, a summary and a discussion are given in Section 5.

2. Methodology

There are different approaches for outlier identification within a dataset of interest using Linear Models (LMs) with n observations (response data), and p independent variables, as given by:

$$Y = \beta_0 + \sum_{j=1}^{p} \beta_j X_j + \epsilon.$$

Here, Y is the response variable and X_j $(j = 1, 2, \ldots, p)$ are the random input variables and $\mathbf{X} = (X_1, X_2, \ldots, X_p)^T$ represents the vector of predictors. The coefficient (or unknown parameter) β_0 is the intercept, the parameters β_j are the regression coefficients related with the independent variables X_j $(j = 1, 2, \ldots, p)$, and ϵ is the random error term, called noise. The data response values and input vectors y_i, $\mathbf{x}_i$ $(i = 1, 2, \ldots, n)$ are inserted into the model, the LM turns into the following linear system:

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}. \tag{1}$$

Here, $\mathbf{y}$ is an $(n \times 1)$-vector of the response variable, $\mathbf{X}$ is a *full rank* $(n \times (p+1))$-matrix of values of explanatory variables with $(1 \times (p+1))$ row vectors $\mathbf{x}_i$ $(i = 1, 2, \ldots, n)$, $\boldsymbol{\beta}$ is a $((p+1) \times 1)$-vector of unknown parameters. Furthermore, $\boldsymbol{\epsilon}$ is an $(n \times 1)$-vector of residuals, regarded to comprise realizations of independent, identically distributed random errors, whose conditional mean and variance are given by $E(\boldsymbol{\epsilon} \mid \mathbf{X}) = 0$ and $\mathrm{Var}(\boldsymbol{\epsilon} \mid \mathbf{X}) = \sigma^2 \mathbf{I}_n$. Here, σ^2 is an unknown parameter and $\mathbf{I}_n$ is the identity matrix of order n. Assuming $n \geqslant p + 1$, the least-squares estimates of $\boldsymbol{\beta}$ and σ^2 are given by $\hat{\boldsymbol{\beta}} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}$ and $\sigma^2 = \mathbf{y}^T(\mathbf{I} - \mathbf{H})\mathbf{y}/(n - p - 1)$, where $\mathbf{H} := \mathbf{X}^T(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T$ is the so-called *hat operator* (Rencher, 2000).

In order to detect outlier(s) for LMs, there are two approaches that use residuals from the robust fit: *direct approaches* and *indirect approaches*. An indirect approach for outlier identification is given through robust regression. The aim of robust regression is to provide stable results in the presence of outliers. Three classes of problems have been addressed with robust regression techniques: problems with outliers in the y-direction (response direction), problems with outliers in the x-space, and problems with outliers in both the y-direction and the x-space. The methods which are most recently used for outlier detection and robust regression are M estimation (Huber, 2009), Least Trimmed Square (LTS) estimation (Rousseeuw and Driessen, 2006) and MSOM (Cook, 1982; Kim *et al.*, 2008). In this paper, we employ the MSOM to identify the outliers in our dataset which we describe next.

2.1. *Mean Shift Outlier Model*

The MSOM is given by:

$$Y = \mathbf{X}^T \boldsymbol{\beta} + \Delta\delta + \epsilon,$$

where $\Delta \in \{0, 1\}$ is a constant selection term, and δ is the unknown parameter for outlier observation. In the presence of an outlier, $\Delta = 1$, and the importance of an outlier are represented by the value $1 \cdot \delta$. The system after all data inserted into the model is as follows:

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{e}_i\delta + \boldsymbol{\epsilon}, \tag{2}$$

where $\mathbf{e}_i$ is the ith unit vector, i.e. $\mathbf{e}_i = (0, \dots, 1, 0, \dots, 0)^T$ $(i = 1, 2, \dots, n)$. In this system, it is assumed that either y_i or $\mathbf{x}_i\boldsymbol{\beta}$ deviates systematically from the model $y_i = \mathbf{x}_i\boldsymbol{\beta} + \epsilon_i$ by some value δ. Then, the ith observation $(y_i, \mathbf{x}_i\boldsymbol{\beta})$ would have a different intercept than the remaining observation, and $(y_i, \mathbf{x}_i\boldsymbol{\beta})$ would hence be an outlier. To check this fact, we test the hypothesis:

$$H_0 : \delta = 0 \quad (\text{i.e.,} E(y) = \mathbf{X}\boldsymbol{\beta}),$$

against the alternative:

$$H_1 : \delta \neq 0 \quad (\text{i.e., } E(y) = \mathbf{X}\boldsymbol{\beta} + \mathbf{e}_i\delta),$$

using the likelihood-ratio test statistic (Rao *et al.*, 1999):

$$F_i = \frac{(RSS(H_0) - RSS(H_1))/1}{RSS(H_1)/(n - p - 1)}. \tag{3}$$

Here, $RSS(H_0)$ is the residual sum of squares in the model $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$, containing all the n observations and $RSS(H_1)$ is the residual sum of squares in the model $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{e}_i\delta + \boldsymbol{\epsilon}$, respectively, i.e.

$$RSS(H_0) = \sum_{i=1}^{n} (y_i - \mathbf{x}_i\boldsymbol{\beta})^2 = \mathbf{y}^T(\mathbf{I} - \mathbf{H})\mathbf{y} = (n - p)\hat{\sigma}^2,$$

$$RSS(H_1) = \sum_{i=1}^{n}(y_i - \mathbf{x}_i\boldsymbol{\beta} - \delta_i)^2,$$

where δ_i will be δ or 0 according to the outlier variable. Also, the relationship between $RSS(H_0)$ and $RSS(H_1)$ can be written as:

$$RSS(H_1) = RSS(H_0) - \frac{\hat{\epsilon}_i^2}{1 - h_{ii}},$$

where $\hat{\boldsymbol{\epsilon}} = (\mathbf{I} - \mathbf{H})\mathbf{y}$, $\hat{\epsilon}_i = \mathbf{e}_i^T\hat{\boldsymbol{\epsilon}}$, and $\mathbf{e}_i^T\mathbf{H}\mathbf{e}_i = h_{ii}$. When the ith observation $(y_i, \mathbf{x}_i)$ is omitted, then estimator of the σ_i^2 is defined by:

$$s_{-i}^2 = \frac{\mathbf{y}_{-i}^T(\mathbf{I} - \mathbf{H}_{-i})\mathbf{y}_{-i}}{n - p - 1},$$

where $\mathbf{H}_{-i}$ and $\mathbf{y}_{-i}$ represent the hat matrix and the response vector after omission of the ith observation, respectively; and σ_i is the standard deviation of the ith residual. If σ_i is taken as $\hat{\sigma}_i = s_{-i}\sqrt{1 - h_{ii}}$, then the test statistic in Eq. (3) may be written as:

$$F_i = \frac{\hat{\epsilon}_i^2}{s_{-i}^2(1 - h_{ii})} = (r_i^*)^2 \quad (i = 1, 2, \ldots, n),$$

where r_i^* is the ith externally Studentized residual (Rao *et al.*, 1999).

For a given dataset of size n with m outliers $(m < n)$ that are detected by direct methods such as the test statistic F_i, Cooks distance or Studentized residuals (Cook, 1982), MSOM can be written as:

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{E}\boldsymbol{\delta} + \boldsymbol{\epsilon}, \tag{4}$$

where $\mathbf{X}$ is a full rank $(n \times (p + 1))$-matrix of explanatory variables, $\mathbf{E}$ is an $(n \times m)$-matrix with m indicator variables, and $\boldsymbol{\delta}$ is an $(m \times 1)$-vector of the regression coefficients of the indicator variables. Then, MSOM can be rewritten as:

$$\mathbf{y} = \mathbf{X}^*\boldsymbol{\beta}^* + \boldsymbol{\epsilon}, \tag{5}$$

where $\mathbf{X}^* = (\mathbf{X} \mid \mathbf{E})$ is an $(n \times (p + 1 + m))$ block matrix constructed by the matrices $\mathbf{X}$ and $\mathbf{E}$, and $\boldsymbol{\beta}^* = (\boldsymbol{\beta}^T, \boldsymbol{\delta}^T)^T$ is an $((p + 1 + m) \times 1)$-vector constructed by the vectors $\boldsymbol{\beta}$ and $\boldsymbol{\delta}$.

Since MSOM (as expressed in Eq. (5)) gives the same residual sum of squares as the model fitted after omitting the relevant observations, it is particularly convenient for studying the regression model in the presence of outliers (Taylan *et al.*, 2014).

2.2. *Improvements on Mean Shift Outlier Model with Conic Multivariate Adaptive Regression Splines*

A major drawback of LM is that in spite of all corrective measures applied, in some cases the constant error variance assumption may not be validated. In such a case, the theory underlying LM fails, and it does not provide the best linear and unbiased estimators for the model parameters anymore. For such datasets, CMARS is known to provide better predictions according to various comparison criteria (Weber *et al.*, 2012; Yerlikaya, 2008). Indeed, CMARS method performs as good as LM does (Yerlikaya-Özkurt *et al.*, 2013) when there are linear relationships between the response variable and the explanatory variables of the given dataset. However, for high-dimensional datasets including a large number of predictors with nonlinear relationships and stochastic dependencies, the complex structure of the data may prevent LM from developing valid and adequate statistical models. For these kinds of estimation problems, numerical evidence indicates that CMARS method provides a better fit than traditional LM (Yerlikaya-Özkurt *et al.*, 2013).

In addition, LMs need human expertise in their use and it may take a longer time to construct parametric models. In certain cases, it may not even be possible to develop LMs. On the other hand, CMARS models are developed automatically and adaptively requiring less human intervention. Finally, for complex datasets, the prediction algorithm should not adapt a parametric form prior to modelling but rather should explore the inherent structure of the dataset to propose a nonparametric form. Thus, for the outlier detection problem, CMARS is employed as a novel and effective tool. Let us consider the following general CMARS model on the relation between input and response:

$$Y = \beta_0 + \sum_{m=1}^{M_{\max}} \beta_m \psi_m\big(\mathbf{x}^m\big) + \epsilon, \tag{6}$$

where Y is a response variable, $\mathbf{x}^m = (x_1^m, x_2^m, \ldots, x_p^m)^T$ is a vector of predictors for the mth basis function and ϵ is an additive stochastic component which is assumed to have zero mean and finite variance. Here, ψ_m $(m = 1, 2, \ldots, M_{\max})$ is the mth basis function, β_m is the unknown coefficient for the mth basis function $(m = 1, 2, \ldots, M_{\max})$ or for the constant 1 $(m = 0)$. The form of the mth basis function is as follows:

$$\psi_m(\mathbf{x}^m) := \prod_{j=1}^{K_m} \big[s_{\kappa_j^m} \cdot \big(x_{\kappa_j^m} - \tau_{\kappa_j^m}\big)\big]_+, \tag{7}$$

where $[q]_+ := \max\{0, q\}$, K_m is the number of truncated linear functions multiplied in the mth basis function, $x_{\kappa_j^m}$ is the input variable corresponding to the jth truncated linear function in the mth basis function, $\tau_{\kappa_j^m}$ is the knot value corresponding to the variable $x_{\kappa_j^m}$, and $s_{\kappa_j^m}$ is the selected sign $+1$ or -1. A special advantage of the form given in Eq. (7) is lying in its ability to estimate the contributions of the basis functions so that both the additive and the interactive effects of the predictors are allowed to determine the dependent variable. *Interaction basis functions* are created by multiplying an existing basis function with a truncated linear function involving a new variable (Friedman, 1991). In a CMARS approximation, both the existing basis functions and the newly created interaction basis functions are used.

The linear system of equations for Eq. (6) is as follows:

$$\mathbf{y} = \boldsymbol{\psi}(\tilde{\mathbf{d}})\boldsymbol{\beta} + \boldsymbol{\epsilon}, \tag{8}$$

where $\boldsymbol{\beta} := (\beta_0, \beta_1, \dots, \beta_{M_{\max}})^T$, $\tilde{\mathbf{d}} := (\mathbf{t}_i^1, \mathbf{t}_i^2, \dots, \mathbf{t}_i^{M_{\max}})^T$ $(i = 1, 2, \dots, n)$ and $\boldsymbol{\psi}(\tilde{\mathbf{d}}) := ((1, \psi_1(\mathbf{t}_1^1), \dots, \psi_{M_{\max}}(\mathbf{t}_1^{M_{\max}}))^T, \dots, (1, \psi_1(\mathbf{t}_n^1), \dots, \psi_{M_{\max}}(\mathbf{t}_n^{M_{\max}}))^T)^T$.

The following model is constructed to employ CMARS for removing the deficiency of MSOM:

$$\mathbf{y} = \boldsymbol{\psi}(\tilde{\mathbf{d}})\boldsymbol{\beta} + \mathbf{e}_i\delta + \boldsymbol{\epsilon}. \tag{9}$$

We then use Penalized Residual Sum of Squares (PRSS) (Hastie *et al.*, 2001) for $M_{\max}$ which is the maximum number of basis functions accumulated in the *forward* stepwise algorithm of CMARS. For the MSOM with CMARS, PRSS has the following form:

$$\begin{aligned} PRSS = {} & \sum_{i=1}^{n} \big(y_i - \boldsymbol{\psi}(\tilde{\mathbf{d}}_i)\boldsymbol{\beta} - \mathbf{e}_i\delta\big)^2 \\ & + \sum_{m=1}^{M_{\max}} \lambda_m \sum_{\substack{|\boldsymbol{\alpha}|=1 \\ \boldsymbol{\alpha}=(\alpha_1,\alpha_2)^T}}^{2} \sum_{\substack{r<s \\ r,s\in V_m}} \int_{Q^m} \beta_m^2 \big[D_{r,s}^{\boldsymbol{\alpha}} \psi_m(\mathbf{t}^m)\big]^2 d\mathbf{t}^m, \end{aligned} \tag{10}$$

where $V_m := \{\kappa_j^m \mid j = 1, 2, \dots, K_m\}$ is the variable set associated with the mth basis function ψ_m, $\mathbf{t}^m = (t_{m_1}, t_{m_2}, \dots, t_{m_{K_m}})^T$ represents the vector of variables which contribute to the mth basis function ψ_m. Furthermore, we refer to

$$D_{r,s}^{\boldsymbol{\alpha}} \psi_m\big(\mathbf{t}^m\big) := \frac{\partial^{|\boldsymbol{\alpha}|}\psi_m}{\partial^{\alpha_1} t_r^m \, \partial^{\alpha_2} t_s^m}\big(\mathbf{t}^m\big)$$

for $\boldsymbol{\alpha} = (\alpha_1, \alpha_2)^T$, $|\boldsymbol{\alpha}| := \alpha_1 + \alpha_2$, where $\alpha_1, \alpha_2 \in \{0, 1\}$. Our optimization problem is based on tradeoff between *accuracy* (i.e. a small residuals) and *complexity*. This tradeoff is established through the penalty parameters λ_m and handled by penalty methods, such as regularization techniques and by conic quadratic programming (Weber *et al.*, 2012). To approximate the multi-dimensional integrals:

$$\int_{Q^m} \beta_m^2 \big[D_{r,s}^{\boldsymbol{\alpha}} \psi_m\big(\mathbf{t}^m\big)\big]^2 d\mathbf{t}^m,$$

a suitable discretization and model approximation are used. In fact, we approximate the discretized form of the integrals by Riemann sums as follows (Yerlikaya, 2008):

$$\begin{aligned}
&\int_{Q^m} \beta_m^2 \big[D_{r,s}^{\boldsymbol{\alpha}} \psi_m(\mathbf{t}^m)\big]^2 d\mathbf{t}^m \\
&\quad \approx \sum_{\sigma^{\kappa_j}} \beta_m^2 \big[D_{r,s}^{\boldsymbol{\alpha}} \psi_m\big(\tilde{t}^m_{i_{\sigma^{\kappa_j}},\kappa_j}, \ldots, \tilde{t}^m_{i_{\sigma^{\kappa_j}},\kappa_j}\big)\big]^2 \prod_{j=1}^{K_m} \big(\tilde{t}^m_{i_{\sigma^{\kappa_j}+1},\kappa_j} - \tilde{t}^m_{i_{\sigma^{\kappa_j}},\kappa_j}\big).
\end{aligned}$$

Here, $(\sigma^{\kappa_j})_{j \in \{1,2,\ldots,p\}} \in \{0, 1, 2, \ldots, n+1\}^{K_m}$. We can rearrange PRSS in this form:

$$\begin{aligned}
PRSS \approx{} & \sum_{i=1}^{n} \big(y_i - \boldsymbol{\psi}(\tilde{\mathbf{d}}_i)\boldsymbol{\beta} - \mathbf{e}_i \delta\big)^2 \\
& + \sum_{m=1}^{M_{\max}} \lambda_m \beta_m^2 \sum_{i=1}^{(n+1)^{K_m}} \left(\sum_{\substack{|\boldsymbol{\alpha}|=1 \\ \boldsymbol{\alpha}=(\alpha_1,\alpha_2)^T}}^{2} \sum_{\substack{r<s \\ r,s \in V_m}} \big[D_{r,s}^{\boldsymbol{\alpha}} \psi_m\big(\hat{\mathbf{t}}_i^m\big)\big]^2 \right) \Delta \hat{\mathbf{t}}_i^m. \qquad (11)
\end{aligned}$$

Moreover, $\hat{\mathbf{t}}_i^m$ and $\Delta \hat{\mathbf{t}}_i^m$ are the notations related with the sequence (σ^{κ_j}):

$$\hat{\mathbf{t}}_i^m = \big(\tilde{t}^m_{i_{\sigma^{\kappa_j}},\kappa_j}, \ldots, \tilde{t}^m_{i_{\sigma^{\kappa_j}},\kappa_j}\big), \qquad \Delta \hat{\mathbf{t}}_i^m = \prod_{j=1}^{K_m} \big(\tilde{t}^m_{i_{\sigma^{\kappa_j}+1},\kappa_j} - \tilde{t}^m_{i_{\sigma^{\kappa_j}},\kappa_j}\big).$$

For a short representation of PRSS, we can rewrite the approximate relation in Eq. (11) as:

$$PRSS \approx \big\| \mathbf{y} - \boldsymbol{\psi}(\tilde{\mathbf{d}})\boldsymbol{\beta} - \mathbf{E}\boldsymbol{\delta} \big\|_2^2 + \sum_{m=1}^{M_{\max}} \lambda_m \sum_{i=1}^{(n+1)^{K_m}} L_{im}^2 \beta_m^2, \qquad (12)$$

where $\mathbf{E}$ is an $n \times m$-matrix with m indicator variables, and $\boldsymbol{\delta}$ is an $m \times 1$-vector of the regression coefficients of the indicator variables. Here, $\| \cdot \|_2$ denotes the Euclidean norm and the numbers L_{im}^2 are defined by their square roots:

$$L_{im} := \left[\left(\sum_{\substack{|\boldsymbol{\alpha}|=1 \\ \boldsymbol{\alpha}=(\alpha_1,\alpha_2)^T}}^{2} \sum_{\substack{r<s \\ r,s \in V_m}} \big[D_{r,s}^{\boldsymbol{\alpha}} \psi_m\big(\hat{\mathbf{t}}_i^m\big)\big]^2 \right) \Delta \hat{\mathbf{t}}_i^m \right]^{1/2}.$$

We consider the approximate formula in Eq. (12) and arrange it as follows, replacing "≈" by "=" from now on:

$$\begin{aligned}
PRSS ={} & \big\| \mathbf{y} - \boldsymbol{\psi}(\tilde{\mathbf{d}})\boldsymbol{\beta} - \mathbf{E}\boldsymbol{\delta} \big\|_2^2 \\
& + \sum_{m=1}^{M_{\max}} \lambda_m \big[(L_{1m}\beta_m)^2 + (L_{2m}\beta_m)^2 + \cdots + (L_{(n+1)^{K_m}m}\beta_m)^2\big],
\end{aligned}$$

$$PRSS = \left\|\mathbf{y} - \boldsymbol{\psi}(\tilde{\mathbf{d}})\boldsymbol{\beta} - \mathbf{E}\boldsymbol{\delta}\right\|_2^2 + \sum_{m=1}^{M_{\max}} \lambda_m \|\mathbf{L}_m \beta_m\|_2^2, \tag{13}$$

where $\mathbf{L}_m = (L_{1m}, L_{2m}, \ldots, L_{(n+1)^{K_m}m})^T$ $(m = 1, 2, \ldots, M_{\max})$. However, rather than a singleton, there is a finite sequence of the *trade-off* or *penalty* parameters $\boldsymbol{\lambda} = (\lambda_1, \lambda_2, \ldots, \lambda_{M_{\max}})^T$ thus this equation is not yet a *Tikhonov regularization problem*. For this reason, let us make a uniform penalization by taking the same λ for each derivative term. Then, our PRRS problem becomes a Tikhonov regularization problem (Aster *et al.*, 2012) with $\lambda > 0$, i.e. $\lambda = \varphi^2$ for some $\varphi \in \mathbb{R}$, as follows:

$$PRSS = \left\|\mathbf{y} - \mathbf{X}^*\boldsymbol{\beta}^*\right\|_2^2 + \lambda\left\|\mathbf{L}^*\boldsymbol{\beta}^*\right\|_2^2. \tag{14}$$

Here, $\mathbf{X}^* = (\boldsymbol{\psi}(\tilde{\mathbf{d}}) \mid \mathbf{E})$ is an $(n \times (M_{\max} + m + 1))$ block matrix constructed by the matrices $\boldsymbol{\psi}(\tilde{\mathbf{d}})$ and $\mathbf{E}$, and $\boldsymbol{\beta}^* = (\boldsymbol{\beta}^T, \boldsymbol{\delta}^T)^T$ is an $((M_{\max} + m + 1) \times 1)$-vector constructed by the vectors $\boldsymbol{\beta}$ and $\boldsymbol{\delta}$. Furthermore, $\boldsymbol{\beta}$ is an $((M_{\max} + 1) \times 1)$-parameter vector to be estimated through the data points (Yerlikaya-Özkurt, 2013).

Let us introduce a block matrix $\mathbf{L}^* = (\mathbf{L} \mid \mathbf{R})$ as follows:

$$\mathbf{L}^* = \begin{bmatrix} \mathbf{L}_{(M_{\max}+1)\times(M_{\max}+1)} & \mathbf{0}_{(M_{\max}+1)\times m} \\ \mathbf{0}_{m\times(M_{\max}+1)} & \mathbf{R}_{m\times m} \end{bmatrix},$$

where $\mathbf{L}$ is a diagonal $(M_{\max} + 1) \times (M_{\max} + 1)$-matrix with the first column $\mathbf{L}_0 = \mathbf{0}_{(n+1)^{K_m}}$ and the other columns being the vectors $\mathbf{L}_m$ $(m = 1, 2, \ldots, M_{\max})$. Moreover, $\mathbf{R}$ is an $m \times m$-matrix whose entries are zeroth-, first- or second-order discrete derivative of $\boldsymbol{\delta}$, the latter two of them complemented with row vectors $\mathbf{0}^T$ to a number of m rows. These derivatives are given by $\boldsymbol{\delta}$ itself, or by first- or second-order difference quotients of $\boldsymbol{\delta}$, respectively (Aster *et al.*, 2012).

Indeed, based on an appropriate choice of a bound $z > 0$, we state the following optimization problem:

$$\begin{aligned} &\underset{\beta^*}{\text{minimize}} \quad \left\|\mathbf{y} - \mathbf{X}^*\boldsymbol{\beta}^*\right\|_2^2 \\ &\text{subject to} \quad \left\|\mathbf{L}^*\boldsymbol{\beta}^*\right\|_2^2 \leqslant z^2. \end{aligned} \tag{15}$$

Let us underline that this choice of z should be the outcome of a careful learning process, based on statistical comparison or performance criteria, with the help of model-free or model-based methods (Weber *et al.*, 2012). In Eq. (15), we have the least-squares objective function $\|\mathbf{y} - \mathbf{X}^*\boldsymbol{\beta}^*\|_2^2$ and the inequality constraint function $-\|\mathbf{L}^*\boldsymbol{\beta}^*\|_2^2 + z^2$ which is requested to be nonnegative for feasibility. Now, we equivalently write our optimization problem as follows (Ben-Tal and Nemirovski, 2001):

$$\begin{aligned}
&\underset{t,\boldsymbol{\beta}^*}{\text{minimize}} \quad t,\\
&\text{subject to} \quad \left\|\mathbf{y}-\mathbf{X}^*\boldsymbol{\beta}^*\right\|_2^2 \leqslant t^2, \quad t \geqslant 0,\\
&\qquad\qquad\quad \left\|\mathbf{L}^*\boldsymbol{\beta}^*\right\|_2^2 \leqslant z^2.
\end{aligned} \tag{16}$$

or, equivalently again,

$$\begin{aligned}
&\underset{t,\boldsymbol{\beta}^*}{\text{minimize}} \quad t,\\
&\text{subject to} \quad \left\|\mathbf{y}-\mathbf{X}^*\boldsymbol{\beta}^*\right\|_2 \leqslant t,\\
&\qquad\qquad\quad \left\|\mathbf{L}^*\boldsymbol{\beta}^*\right\|_2 \leqslant z.
\end{aligned} \tag{17}$$

We use modern methods of *convex optimization techniques*, especially, from Conic Quadratic Programming (CQP) where we employ the basic notation (Nesterov and Nemirovski, 1994):

$$\begin{aligned}
&\underset{t,\boldsymbol{\beta}^*}{\text{minimize}} \quad t,\\
&\text{such that}\\
&\boldsymbol{\chi} := \begin{bmatrix} \mathbf{0}_{n\times 1} & \mathbf{X}^* \\ 1 & \mathbf{0}^T_{(M_{\max}+m+1)\times 1} \end{bmatrix} + \begin{bmatrix} t \\ \boldsymbol{\beta}^* \end{bmatrix} + \begin{bmatrix} -\mathbf{y} \\ 0 \end{bmatrix},\\
&\boldsymbol{\eta} := \begin{bmatrix} \mathbf{0}_{(M_{\max}+m+1)\times 1} & \mathbf{L}^* \\ 0 & \mathbf{0}^T_{(M_{\max}+m+1)\times 1} \end{bmatrix} \begin{bmatrix} t \\ \boldsymbol{\beta}^* \end{bmatrix} + \begin{bmatrix} \mathbf{0}_{(M_{\max}+m+1)\times 1} \\ z \end{bmatrix},\\
&\boldsymbol{\chi} \in \mathbf{L}^{n+1}, \quad \boldsymbol{\eta} \in \mathbf{L}^{M_{\max}+m+2},
\end{aligned}$$

where L^{n+1}, $L^{M_{\max}+m+2}$ are the $(n+1)$- and $(M_{\max}+m+2)$-dimensional *second-order* (or *Lorentz*) *cones*. A primal-dual optimal solution is $(t, \boldsymbol{\theta}, \boldsymbol{\chi}, \boldsymbol{\eta}, \boldsymbol{\omega}_1, \boldsymbol{\omega}_2)$ (Weber *et al.*, 2012; Yerlikaya-Özkurt, 2013).

In this study, we are addressing a need for regularization and perform it in two ways under different aspects (Yerlikaya-Özkurt, 2013):

1. We take into account possible outliers by identifying them through mathematical modelling in terms of their impact on the model. This is a first way to make the model "regular" with respect to the existence of outliers. By the basic settings of our model, we "reach out" to possible outliers and represent their possible character and contribution to the model by a parameter δ, which has to be assessed numerically.
2. In both a standard MSOM and a nonlinear CMARS model, we perform Tikhonov regularization for addressing both model accuracy and complexity, too.

3. Dataset

We employed the recently compiled Turkish strong ground motion data (Akkar *et al.*, 2010; Gülerce *et al.*, 2013) in this study. The dataset includes 290 strong ground motion

Table 1
Number of records for different magnitude bins.

	$M_w \geqslant 7.0$	$6.0 \leqslant M_w < 7.0$	$5.0 \leqslant M_w < 6.0$
# records	28	25	237

Table 2
NEHRP site class definitions and number of records per each site class for the given dataset.

NEHRP site class	Soil profile name	Shear wave velocity (m/s)	# Records for given dataset
A	Hard rock	$V_{s30} > 1524$	0
B	Rock	$762 < V_{s30} \leqslant 1524$	0
C	Very dense soil soft rock	$366 < V_{s30} \leqslant 762$	172
D	Stiff soil	$183 < V_{s30} \leqslant 366$	117
E	Soft soil	$V_{s30} < 183$	1

records with a moment magnitude range of $5.0 \leqslant M_w \leqslant 7.6$ and Joyner–Boore distance (R_{jb}) range of $0 < R_{jb} < 200$ km. In this study, we employed the processed dataset as presented in Yerlikaya-Özkurt *et al.* (2014), while the raw version of the data can be found on the web page `http://daphne.deprem.gov.tr:89/` operated by the Earthquake Department of the Turkish Disaster and Emergency Management Agency. To identify the local soil conditions at the stations, consistent with the current literature, we use the 30 m average shear wave velocity (V_{s30}) as a direct measure in this study.

Table 1 displays the number of records in terms of different magnitude bins. Table 2 shows the different site classes with their definitions and number of records for each site class in the dataset employed herein (the NEHRP site classes are taken from the site classifications available in Section 1613.3.2 "Site Class Definitions" of the International Building Code, published in 2012 by the International Code Council). Finally, Fig. 1 displays the PGA-distance distribution of the dataset with respect to different soil conditions for different magnitude bins.

In this study only the earthquakes with a strike-slip fault mechanism, which is the major faulting style on the North Anatolian Fault zone in Turkey, are used.

4. Application

In order to find the potential outliers for each magnitude bin, we apply the following outlier detection procedure (Montgomery and Peck, 1992; Hadi and Simonoff, 1993):

Step 1: Multivariate linear model is constructed to fit the data.

Step 2: The fit values and ordinary residuals are computed by using the model from Step 1.

Step 3: Studentized residual, leverage, measure of influence such as Cooks distance and measure of model performance such as scaled change in regression coefficients, scaled change in fitted values, and change in covariance are computed. The definitions can be found in Montgomery and Peck (1992).

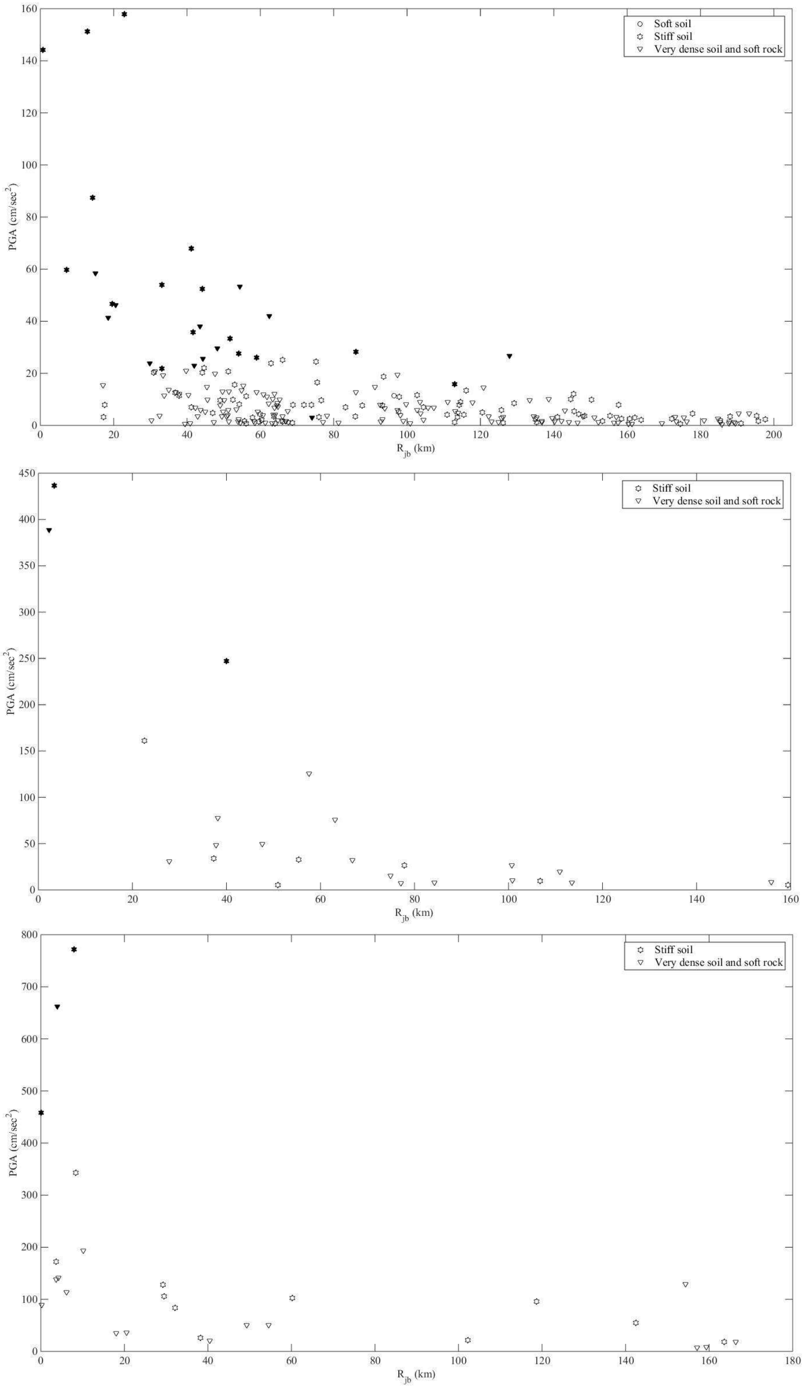

Fig. 1. Ground motion dataset in terms of different magnitude bins (a) $5.0 \leqslant M_w < 6.0$, (b) $6.0 \leqslant M_w < 7.0$, (c) $M_w \geqslant 7.0$. Each soil condition is represented with a different data marker in the legend. Potential outliers are represented by filling in the markers.

Table 3
Number of potential outliers for each bin.

	$M_w \geqslant 7.0$	$6.0 \leqslant M_w < 7.0$	$5.0 \leqslant M_w < 6.0$
# potential outliers	3	3	29

Step 4: The potential outlier is removed from the dataset and Steps 1 and 2 are repeated to check for a better fit. (An observation is a potential outlier if the fit without that observation is better than the fit including the observation.)
Step 5: In order to find and remove the other potential outlier observations, Steps 1–4 are repeated until all of the outlier observations are identified.

After applying these steps to each magnitude bin, we obtain the number of potential outliers for each bin as presented in Table 3. We note that potential outliers for each magnitude bin are studied one by one and they are confirmed to be out-of-range values given the magnitude, distance and soil conditions. When Fig. 1 is studied carefully, it is possible to see that around the same/similar magnitude and same/similar distance (R_{jb}) levels within each M_w bin; some PGA values corresponding to stiff soil conditions are smaller than PGA values corresponding to "very dense soil and soft rock" conditions. This is physically not expected as the stiff soil conditions generally have larger amplifications than very dense soil and soft rock. Thus, these data are interpreted as the potential outliers. It must be stated that some observations in our dataset might not be correctly measured due to potential instrumental failures during earthquake shakings. Such erroneous data points were removed prior to the application of the outlier detection algorithm.

After the detection of potential outliers, we constructed three alternative models on our standardized dataset (with logarithmic response) which are namely: MSOM, CMARS and MSOM-CMARS. In our application, MARS basis functions are built by using the Salford MARS software program; indeed, the R package "Earth" can also be used for the construction of basis functions (Milborrow, 2009). The CMARS model parameters are constructed by running a MATLAB® code via the optimization software MOSEK™. In the near future, CMARS algorithm will be provided in R program to the interested researchers.

We note that these models are constructed on the complete dataset including the previously identified outliers. To evaluate the comparative efficiency of these models, we compute a set of performance measures. In addition, we place two well-known and widely accepted parametric GMPEs to compare the non-parametric GMPEs we propose in this paper. The selected parametric GMPEs are the predictive equations by Boore and Atkinson (2008), Akkar and Çağnan (2010). These models are represented as BA08 and AC10, respectively, from this point onward.

Through a comparison of performance measures such as Mean Absolute Error (MAE), Mean Squared Error (MSE), Root Mean Squared Error (RMSE), Correlation Coefficient (r), Multiple Coefficient of Determination (R^2), Adjusted R^2 (Adj-R^2), and Mean Absolute Percentage Error (MAPE) in Table 4, non-parametric model (CMARS) performs better than the parametric ground motion prediction models (BA08, AC10 and MSOM). On the other hand, among the parametric models, MSOM shows a better performance than

Table 4
Performance results of the models built for the given dataset.

Performance measure	BA08	AC10	MSOM	CMARS	MSOM-CMARS
MAE	1.0037	0.6725	0.2345	0.2364	0.1854*
MSE	1.6230	0.7527	0.1098	0.0997	0.0797*
RMSE	1.2740	0.8676	0.3313	0.3158	0.2823*
r	0.8020	0.8216	0.8803	0.8856	0.9222*
R^2	0.6433	0.6751	0.7750	0.7842	0.8504*
Adj-R^2	0.6383	0.6717	0.7410	0.7647	0.8120*
MAPE	114.2122	50.1166	47.2743	31.9453	30.4293*

*Indicates better performance.

BA08 and AC10, since it considers the outliers in the dataset. Finally, MSOM-CMARS is observed to yield the smallest misfits among all models. This is because this hybrid model combines the power of CMARS in capturing the data structure with the effective outlier modelling of MSOM.

It is important to verify that there is no systematic bias in model residuals with respect to each independent variable M_w, R_{jb} and V_{s30}.

In the model form of MSOM-CMARS presented in Appendix A, it is observed that all basis functions for this dataset have a main effect (without any interaction effects). We also observe that the coefficients of the potential outliers are smaller than those of basis functions indicating a smaller influence of the outliers on the final model. Therefore, we construct a model which takes into account by quantifying the effects of outliers without eliminating them.

5. Summary and Discussion

The fundamental objective of this study was to develop a robust computational method for the data prediction problem with the help of convex optimization within the existence of outliers in the dataset. For this purpose, we proposed a hybrid approach that takes advantages of CMARS, MSOM and CQP. The motivation for using a CQP is due to appealing properties of the model and fast algorithms that are available to solve such a model. Specifically, the set of feasible solutions of the CQP problem is convex, which guarantees convergence to a globally optimal solution.

To assess the performance of the proposed approach, we compared our results against other models. The results indicated that since the optimal levels of process parameters yield desired responses in the application, the hybrid MSOM-CMARS model performed the best among its current alternatives that include parametric models and also a non-parametric model, CMARS.

The proposed study is a novel approach to handle outliers within the ground motion prediction framework in a systematic and effective way. A major advantage of this model over other available robust estimation algorithms is the non-existence of assumptions that should be validated for effective modelling in the existence of outliers.

It is always possible to improve the MSOM-CMARS models in future applications. Several suggestions can be summarized as follows:

- In this study, we worked on a relatively limited ground motion dataset in terms of total number of records per each magnitude bin and site class. Since CMARS works effectively for high-dimensional data, for more complete ground motion datasets the results would indicate an even better performance compared to other models that do not handle outliers.
- As the studied dataset (Turkish ground motion database) expands with records from future earthquakes, it is possible to test the proposed model for the prediction of the newly added data. It is also possible to predict the anticipated ground motion levels in potential scenario earthquakes whenever necessary.

Acknowledgements. The first author, F. Yerlikaya-Özkurt, was supported by the TUBITAK Domestic Doctoral Scholarship Program.

Appendix A

The model form of MSOM-CMARS for given dataset is presented next:

$$\begin{aligned} Y_{PGA} = {} & -0.3371 - 13.6579 \cdot \psi_1 + 6.2582 \cdot \psi_2 - 9.3666 \cdot \psi_3 - 4.6509 \cdot \psi_4 \\ & - 21.15875 \cdot \psi_5 + 14.3226 \cdot \psi_6 + 49.7243 \cdot \psi_7 - 3.9772 \cdot \psi_8 \\ & - 42.0155 \cdot \psi_9 + 2.6307 \cdot \psi_{10} - 2.6703 \cdot \psi_{11} + 2.5886 \cdot \psi_{12} \\ & + 2.36177 \cdot \psi_{13} + 9.6167 \cdot \psi_{14} + 5.8253 \cdot \psi_{15} - 22.3286 \cdot \psi_{16} \\ & + 10.9114 \cdot \psi_{17} + 16.6962 \cdot \psi_{18} - 15.9818 \cdot \psi_{19} + 12.6404 \cdot \psi_{20} \\ & + 10.8503 \cdot \psi_{21} - 9.3256 \cdot \psi_{22} + 6.2780 \cdot \psi_{23} \\ & - 1.3084 \cdot \psi_{24} + 0.3408 \cdot e_1 + 0.8214 \cdot e_2 + 0.6297 \cdot e_3 + 0.4141 \cdot e_4 \\ & + 0.4257 \cdot e_5 + 0.6856 \cdot e_6 + 0.5296 \cdot e_7 + 0.6834 \cdot e_8 + 0.8725 \cdot e_9 \\ & + 1.0976 \cdot e_{10} + 0.4252 \cdot e_{11} + 0.3065 \cdot e_{12} + 0.5124 \cdot e_{13} + 0.6292 \cdot e_{14} \\ & + 0.6182 \cdot e_{15} + 0.8642 \cdot e_{16} + 0.5494 \cdot e_{17} + 0.9180 \cdot e_{18} + 0.4899 \cdot e_{19} \\ & + 0.3350 \cdot e_{20} + 0.4296 \cdot e_{21} + 0.7997 \cdot e_{22} + 0.4787 \cdot e_{23} + 0.3114 \cdot e_{24} \\ & + 0.1509 \cdot e_{25} + 0.3048 \cdot e_{26} + 0.5317 \cdot e_{27} + 0.8511 \cdot e_{28} + 0.7057 \cdot e_{29} \\ & + 0.3488 \cdot e_{30} + 0.5918 \cdot e_{31} + 0.4542 \cdot e_{32} + 0.2671 \cdot e_{33} + 0.7499 \cdot e_{34} \\ & - 0.0946 \cdot e_{35} + \epsilon, \end{aligned}$$

where Y_{PGA} is the PGA prediction of CMARS model. The corresponding basis functions ($\psi_m = \psi_m(\mathbf{x}^m)$ $(m = 1, 2, \ldots, 24)$) are given as follows:

$$\begin{aligned}
\psi_1 &= \max\{0, R_{jb} - 0.2540\}, & \psi_2 &= \max\{0, 0.2540 - R_{jb}\},\\
\psi_3 &= \max\{0, M_w - 0.0654\}, & \psi_4 &= \max\{0, V_{s30} - 0.8512\},\\
\psi_5 &= \max\{0, R_{jb} - 0.4945\}, & \psi_6 &= \max\{0, R_{jb} - 0.3486\},\\
\psi_7 &= \max\{0, V_{s30} - 0.4213\}, & \psi_8 &= \max\{0, V_{s30} - 0.5345\},\\
\psi_9 &= \max\{0, V_{s30} - 0.4073\}, & \psi_{10} &= \max\{0, V_{s30} - 0.7341\},\\
\psi_{11} &= \max\{0, M_w - 0.4615\}, & \psi_{12} &= \max\{0, M_w - 0.8077\},\\
\psi_{13} &= \max\{0, M_w - 0.2308\}, & \psi_{14} &= \max\{0, M_w - 0.0192\},\\
\psi_{15} &= \max\{0, R_{jb} - 0.0936\}, & \psi_{16} &= \max\{0, V_{s30} - 0.4613\},\\
\psi_{17} &= \max\{0, V_{s30} - 0.4896\}, & \psi_{18} &= \max\{0, R_{jb} - 0.5101\},\\
\psi_{19} &= \max\{0, R_{jb} - 0.3183\}, & \psi_{20} &= \max\{0, R_{jb} - 0.2841\},\\
\psi_{21} &= \max\{0, V_{s30} - 0.3717\}, & \psi_{22} &= \max\{0, V_{s30} - 0.2925\},\\
\psi_{23} &= \max\{0, V_{s30} - 0.2511\}, & \psi_{24} &= \max\{0, V_{s30} - 0.2056\}.
\end{aligned}$$

References

Akkar, S., Çağnan, Z. (2010). A local ground-motion predictive model for Turkey, and its comparison with other regional and global ground-motion models. *Bulletin of the Seismological Society of America*, 100(6), 2978–2995.

Akkar, S., Çağnan, Z., Yenier, E., Erdoğan, O., Sandıkkaya, M.A., Gülkan, P. (2010). The recently compiled Turkish strong motion database: preliminary investigation for seismological parameters. *Journal of Seismology*, 14(3), 457–479.

Alavi, A.H., Gandomi, A.H. (2011). Prediction of principal ground-motion parameters using a hybrid method coupling artificial neural networks and simulated annealing. *Computers and Structures*, 89, 2176–2194.

Aster, R.C., Borchers, B., Thurber, C. (2012). *Parameter Estimation and Inverse Problems*. Academic Press, Burlington.

Barnett, V., Lewis, T. (1994). *Outliers in Statistical Data*. Wiley, Great Britain.

Ben-Tal, A., and Nemirovski, A. (2001). *Lectures on Modern Convex Optimization: Analysis, Algorithms and Engineering Applications*. SIAM, Philadelphia.

Boore, D.M., Atkinson, G.M. (2008). Ground-motion prediction equations for the average horizontal component of PGA, PGV, and 5%-damped PSA at spectral periods between 0.01 s and 10.0 s. *Earthquake Spectra*, 24(1), 99–138.

Calvin, J.M., Žilinskas, A. (2005). One-dimensional global optimization for observations with noise. *Computers and Mathematics with Applications*, 50, 157–169.

Cook, R.D. (1982). *Residuals and Influence in Regression*. Chapman and Hall, New York.

Friedman, J.H. (1991). Multivariate adaptive regression splines. *The Annals of Statistics*, 19(1), 1–141.

Gülerce, Z., Kargıoğlu, B., Abrahamson, N.A. (2013). Turkey-adjusted NGA-W1 horizontal ground motion prediction models. *Earthquake Spectra*, submitted for publication.

Hadi, A.S., Simonoff, J.S. (1993). Procedures for the Identification of multiple outliers in linear models. *Journal of the American Statistical Association*, 88(424), 1264–1272.

Hastie, T., Tibshirani, R., Friedman, J.H. (2001). *The Elements of Statistical Learning*. Springer Verlag, New York.

Huber, P.J. (2009). *Robust Statistics*. John Wiley & Sons, Hoboken, NJ.

Kim, S.S., Park, S.H., Krzanowski, W.J. (2008). Simultaneous variable selection and outlier identification in linear regression using the mean-shift outlier model. *Journal of Applied Statistics*, 35(3), 283–291.

Lane, K. (2002). What is robust regression and how do you do it? *Paper presented at the Annual Meeting of the Southwest Educational Research Association*.

Milborrow, S. (2009). Earth: multivariate adaptive regression spline models.

Montgomery, D.C., Peck, E.A. (1992). *Introduction to Linear Regression Analysis*. John Wiley & Sons, New York.

Nesterov, Y.E., Nemirovski, A.S. (1994). *Interior-Point Polynomial Algorithms in Convex Programming*. SIAM, Philadelphia.

Peruš, I., Fajfar, P. (2010). Ground-motion prediction by a non-parametric approach. *Earthquake Engineering and Structural Dynamics*, 39(12), 1395–1416.

Rao, C.R., Toutenburg, H., Fieger, A. (1999). *Linear Models: Least Squares and Alternatives*. Springer.

Rencher, A.C. (2000). *Linear Models in Statistics*. John Wiley & Sons, New York.

Rousseeuw, P.J., Driessen, K.V. (2006). Computing LTS regression for large data sets. *Data Mining and Knowledge Discovery*, 12, 29–45.

Rousseeuw, P.J., Leroy, A.M. (1987). *Robust Regression and Outlier Detection*. John Wiley & Sons, New York.

Spudich, P., Joyner, W.B., Lindh, A.G., Boore, D.M., Margaris, B.M., Fletcher, J.B. (1999). SEA99: a revised ground motion prediction relation for use in extensional tectonic regimes. *Bulletin of the Seismological Society of America*, 89(5), 1156–1170.

Taylan, P., Yerlikaya-Özkurt, F., Weber, G.W. (2014). An approach to mean shift outlier model (MSOM) by Tikhonov regularization and conic programming. *Intelligent Data Analysis*, 18(1), 79–94.

Tezcan, J., Cheng, Q. (2012). Support vector regression for estimating earthquake response spectra. *Bulletin of Earthquake Engineering*, 10(4), 1205–1219.

Weber, G.W., Batmaz, İ., Köksal, G., Taylan, P., Yerlikaya-Özkurt, F. (2012). CMARS: a new contribution to nonparametric regression with multivariate adaptive regression splines supported by continuous optimization. *Inverse Problems in Science and Engineering*, 20(3), 371–400.

Yerlikaya, F. (2008). *A new contribution to nonlinear robust regression and classification with MARS and its application to data mining for quality control in manufacturing*. MSc Thesis, Institute of Applied Mathematics, Middle East Technical University, Ankara, Turkey.

Yerlikaya-Özkurt, F. (2013). *Refinements, extensions and modern applications of conic multivariate adaptive regression splines*. PhD Thesis, Institute of Applied Mathematics, Middle East Technical University, Ankara, Turkey.

Yerlikaya-Özkurt, F., Askan, A., Weber, G.W. (2014). An alternative approach to ground motion prediction problem by a non-parametric adaptive regression method. *Engineering Optimization*, 46(12), 1651–1668.

Yerlikaya-Özkurt, F., Batmaz, İ., Weber, G.W. (2013). A review and new contribution on conic multivariate adaptive regression splines (CMARS): a powerful tool for predictive data mining. In: Zilberman, D., Pinto, A. (Eds.), *Springer Volume Modeling, Optimization, Dynamics and Bioeconomy, series Springer Proceedings in Mathematics*.

Žilinskas, A. (2010). On similarities between two models of global optimization: statistical model and radial basis functions. *Journal of Global Optimization*, 48, 173–182.

F. Yerlikaya-Özkurt received her Master's and Doctoral degrees in scientific computing at Institute of Applied Mathematics, Middle East Technical University in 2008 and 2013, respectively. At present she is postdoc researcher in the Department of Industrial and Systems Engineering at Lehigh University. Her main research interests include convex optimization, data mining and computational statistics.

A. Askan is an associate professor of civil engineering at Middle East Technical University. She obtained her doctoral degree from Carnegie Mellon University in 2006. Her research focuses on computational earthquake engineering and engineering seismology with emphasis on numerical methods. She is also affiliated with Earthquake Studies department and Institute of Applied Mathematics at Middle East Technical University.

G.-W. Weber is a professor at Institute of Applied Mathematics, Middle East Technical University, Ankara, Turkey. His research is on optimization and control, OR, finance, life and human sciences, data mining and inverse problems. He received diploma and doctoral degree in mathematics and economics/BA at RWTH Aachen, and his habilitation at TU Darmstadt. He held professorships by proxy at University of Cologne, and TU Chemitz (Germany); he has international affiliations and honorary positions, and is "Advisor to EURO Conference".

Iškilios optimizacijos pagrindu sudarytas hibridinis skaitmeninis metodas, skirtas išskirčių problemoms: taikymai prognozuoti žemės drebėjimų sukeliamam grunto judėjimui

Fatma YERLİKAYA-ÖZKURT, Aysegul ASKAN, Gerhard-Wilhelm WEBER

Daugelis praktinių prognozavimo uždavinių sprendžiami statistinio modeliavimo metodu. Tačiau prognostiniai modeliai gali duoti klaidingų rezultatų esant išskirtims. Daugelyje praktinių uždavinių svarbu neeliminuoti išskirčių – jas identifikuoti ir nagrinėti. Tam ir skirtas hibridinis metodas, pagrįstas kūginiu kvadratiniu programavimu, kuris pritaikytas nagrinėti žemės drebėjimo duomenims. Šiuo metodu siekiama minimizuoti išskirčių įtaką regresijos įverčiams, taip pat atsižvelgti į duomenų netiesiškumus. Gauti rezultatai palyginti su rezultatais, gautais plačiai taikomais parametriniais ir neparametriniais žemės drebėjimo prognozavimo metodais.