Hindawi
Complexity
Volume 2019, Article ID 2360610, 20 pages
https://doi.org/10.1155/2019/2360610

Research Article

Public Procurement Announcements in Spain: Regulations, Data Analysis, and Award Price Estimator Using Machine Learning

Manuel J. García Rodríguez, Vicente Rodríguez Montequín, Francisco Ortega Fernández, and Joaquín M. Villanueva Balsera

Project Engineering Area, University of Oviedo, Oviedo 33004, Spain

Correspondence should be addressed to Vicente Rodríguez Montequín; montequi@uniovi.es

Received 27 June 2019; Revised 13 September 2019; Accepted 27 September 2019; Published 14 November 2019

Guest Editor: Benjamin M. Tabak

The largest project managers and adjudicators of a country, both by number of projects and by cost, are public procurement agencies. Therefore, knowing and characterising public procurement announcements (tenders) is fundamental for managing public resources well. This article presents the case of public procurement in Spain, analysing a dataset from 2012 to 2018: 58,337 tenders with a cost of 31,426 million euros. Many studies of public procurement have been conducted globally or theoretically, but there is a dearth of data analysis, especially regarding Spain. A quantitative, graphical, and statistical description of the dataset is presented. Mainly, the analysis is of the relation between the award price and the bidding price. An award price estimator is proposed that uses the random forest regression method. A good estimator would be very useful and valuable for companies and public procurement agencies. It would be a key tool in their project management decision making. Finally, a similar analysis, employing a dataset from European countries, is presented to compare and generalise the results and conclusions. Hence, this is a novel study which fills a gap in the literature.

1. Introduction

Every year, public authorities in European countries spend around 14% of GDP on public procurement, about 1.9 trillion euros [1], which is the latest estimate (2017) not including spending by utility companies. Spain is also similar, which spends around 10% to 20% of GDP [2]. Public procurement is very important in sectors such as civil construction, energy, transport, defence, IT, or health services. Therefore, it is crucial to analyse the public procurement notices, also called requests for tenders or simply tenders, to understand their behaviour in terms of prices, bidding companies, duration of projects, types of work, etc.

The growing awareness of public procurement as an innovative policy tool has recently sparked the interest of both policy makers and researchers [3]. The open data associated with public procurement and other open government data initiatives [4] are increasing mainly due to the following factors:

(i) Technological factors: software tools to manipulate big data and machine learning algorithms to analyse data (e.g., to make predictions) [5, 6].

(ii) Bureaucratic factors: standardisation of contracting language e-procurement [7, 8] and the benefits of the digitalisation of public procurement agencies [9].

(iii) Political factors: greater transparency in political decision making and design of methods of selecting suppliers for public procurement [10].

(iv) Economic factors: accuracy of the estimation of the cost [11], contract renegotiation [12], risk and uncertainty in the contracts [13], estimation of bidder participation in tenders [14] and its impact on prices [15], and globalisation—companies competing in markets far away from their origin [1].

(v) Social factors: less tolerance for inefficient political management or political irregularities in the procedure [16] and greater transparency and flexibility

in award mechanisms between public procurement agencies and private companies [17].

The layout of this paper is connected with the method employed in the research, as depicted in Figure 1. Section 2 summarises the legislation regarding public procurement notices. A tender is organised in fields, but nevertheless, it is necessary to preprocess the information to produce the dataset. The data fields involved in the process as well as how the data are preprocessed are described. Section 3 analyses the dataset (main characteristic values, correlation, dispersion, etc.), lists the evaluation metrics used (types of errors), and makes a quantitative and graphical analysis of two fundamental fields: the tender price and the award price. The competition in public tenders and its impact on savings have been analysed: how the award price is affected by the competitiveness of the companies. In Section 4, an estimator of the award price is proposed using the machine learning algorithm random forest for regression. Several fields of the tender (the name of the public procurement agency, type of contract, geographical location, type of work or service, duration, date, etc.) have been used to make the prediction. The success of the estimator is analysed based on the evaluation metrics defined previously. Furthermore, a similar analysis employing a dataset from other European countries is presented. Lastly, some concluding remarks and avenues for future research are presented in Section 5.

As far as we know, this article is the first attempt to provide an award price estimator for all types of tenders in a country using machine learning algorithms. Similar articles dealing with this topic [18, 19] have been published recently but only for construction projects and small datasets. It is typical to find literature only applied to construction projects; this is mainly because they are the biggest public procurement projects. On the contrary, the approach of this article is from a multidisciplinary perspective, and it analyses a large volume of data using machine learning techniques.

2. Spanish Public Tenders (2012–2018): Description of the Dataset

In this section, the origin and nature of the Spanish public procurement processes are analysed. Section 2.1 presents a summary of the legislation associated with public procurement and the reuse of public information. Section 2.2 lists the fields of the public procurement notice with information that appears in the announcement. Section 2.3 explains how the original information has been preprocessed to finally obtain a dataset which is valid for statistical and mathematical analysis.

2.1. European and Spanish Legislation on Public Procurement and on the Reuse of Public Information. At the European and Spanish levels, laws have been developed related to the reuse of public sector information and procurement or contracting in the public sector. They are summarised in Table 1. According to *Spanish Law 20/2013*, the website of the Public Sector Contracting Platform (P.S.C.P.) of Spain has to publish the public procurement notices and their resolutions of all contracting agencies belonging to the Spanish Public Sector.

With regard to official announcements of Spanish tenders outside Spain, Article 135 of *Law 9/2017* establishes that when tenders are subject to harmonised regulations (those with an amount greater than a threshold or with certain characteristics, stipulated in Articles 19 to 23), tenders have to also be published in *The Official Journal of the European Union* (OJEU) [20]. When the public contracting authority considers it appropriate, tenders not subject to harmonised regulations can be announced in the OJEU. The Europe Union (EU) has an Open Data Portal [21] which was set up in 2012, following *Commission Decision 2011/833/EU* on the reuse of commission documents. All EU institutions are invited to make their data publicly available whenever possible.

Furthermore, there is a portal called Tenders Electronic Daily (TED) [22] dedicated to European public procurement. It provides free access to business opportunities in the EU, the European Economic Area, and beyond.

2.2. Data Fields of Spanish Public Procurement Notices. The information of public procurement notices is defined in *Spanish Law 9/2017*, Annex III "Information that has to appear in the announcements." P.S.C.P. has an open data section for the reuse of this information (in compliance with the publicity obligations established in *Law 9/2017*) which will be used in this article to generate the dataset. The information is provided by the Ministry of Finance (link in the Data Availability section) and has been published as open data since 2012 and updated monthly in XML format.

The fields of the public procurement notices are numerous, and they can completely define the tender. The most important fields are as follows (more details in Table 2):

(i) Announcement fields: tender status, contract file number, object of the contract, tender price (budget), duration of the contract, CPV classification, contract type, contract subtype, place of execution, lots, type of procedure, contracting system, type of processing, contracting body, place and deadline for submission of tenders, participation requirements, award criteria, subcontracting conditions, contract modifications, etc.

(ii) Award fields: award result, identity of the winning company (CIF and company name), award price, number of received offers, maximum and minimum received bids, etc.

Not all fields have been selected (last column in Table 2) to mathematically analyse the tenders for several reasons:

(1) Some fields are usually empty or have inconsistent data or errors.

(2) Not all fields have the same importance. For example, the tender price is more important than the language of the tender document.

(3) The content of many of these fields is textual, which makes their mathematical modelling very complex.

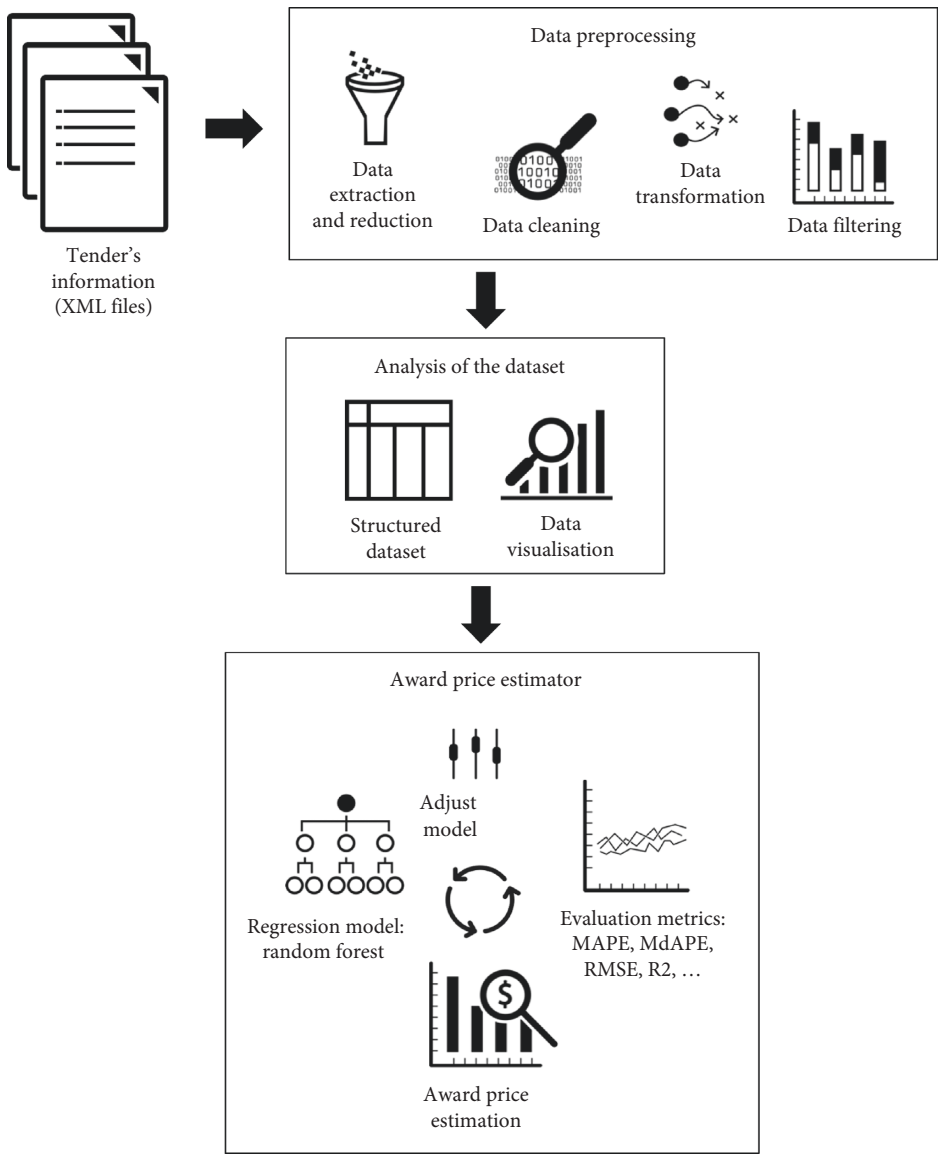


Figure 1: Flowchart of the data analysis and award price estimator.

2.3. Data Preprocessing. It is necessary to carry out several steps to preprocess the data. This is a laborious task because the tender's information has not been verified automatically to correct human errors. The preprocessing can be divided into the following 5 consecutive tasks:

(1) *Data Extraction*. Structured data are stored in text files (XML format). A script has been created to read the fields recursively, saving in the database one tender per row and as many columns as there are fields to be stored.

(2) *Data Reduction*. Around 60 fields are selected; a priori they are interesting for the performance of a statistical and mathematical analysis.

(3) *Data Cleaning*. The data are cleaned. For example, deleting spaces, punctuation marks, and special characters, conversion to capital letters, deleting data with fixed structure (postal code, CPV, CIF, etc.) which do not obey the structure's rules, etc.

(4) *Data Transformation*. Basically, four types of transformations are carried out:

   (a) *Normalisation*. This consists of homogenising the fields. For example, converting dates to time stamps.

   (b) *Aggregation*. This consists of adding new useful fields for the analysis. For example, creating a new field which is the first two numbers of the CPV classification (common procurement vocabulary).

   (c) *Data Enhancement*. It serves to create fields with external information and thus enables checking the consistency of the extracted data. For example, employing the postal code of the tender, it has generated its geographical

Table 1: Laws about public procurement and the reuse of public sector information.

Law	Description	Level	Permanent link
Directive 2003/98/EC	Reuse of public sector information	Europe	http://data.europa.eu/eli/dir/2003/98/oj
Directive 2013/37/EU	Modifying previous directive 2003/98/EC	Europe	http://data.europa.eu/eli/dir/2013/37/oj
Directive 2007/2/EC	Establishing an Infrastructure for Spatial Information in the European Community (INSPIRE)	Europe	http://data.europa.eu/eli/dir/2007/2/oj
Law 37/2007	Transposing into Spanish law the European directive 2003/98/EC	Spain	https://boe.es/eli/es/l/2007/11/16/37
Royal Decree-Law 1495/2011	Developing the Spanish law 37/2007	Spain	https://boe.es/eli/es/rd/2011/10/24/1495
Commission Decision 2011/833/EU	On the reuse of commission documents	Europe	http://data.europa.eu/eli/dec/2011/833
Law 19/2013	Transparency, access to public sector information and good governance	Spain	https://boe.es/eli/es/l/2013/12/09/19
Law 20/2013	Market unit guarantee	Spain	https://boe.es/eli/es/l/2013/12/09/20
Law 18/2015	Transposing into Spanish law the European directive 2013/37/EU	Spain	https://boe.es/eli/es/l/2015/07/09/18
Directive 2014/23/EU	Award of concession contracts	Europe	http://data.europa.eu/eli/dir/2014/23/oj
Directive 2014/24/EU	Public procurement	Europe	http://data.europa.eu/eli/dir/2014/24/oj
Law 9/2017	Transposing into Spanish law the previous European directives 2014/23/UE and 2014/24/UE	Spain	https://boe.es/eli/es/l/2017/11/08/9

Table 2: Most relevant data fields in the public procurement notices (tenders) used in the dataset

Name	Description	Name column dataset
Tender status	Status of the tender during the development of the procedure: prior notice, in time, pending adjudication, awarded, resolved or cancelled	Not used (similar to Result_code)
Contract file number	Unique identifier for a contract file	Not used
Object of the contract	Summary description of the contract	Not used (unstructured textual information)
Public procurement agency	Public procurement agency that made the tender: name, identifier (NIF or DIR3), website, address, postal code, city, country, contact name, telephone, fax, e-mail, etc	**Name_Organisation Postalzone Postalzone_CCAA Postalzone_Province Postalzone_Municipality**
Tender price	Amount of bidding budgeted (taxes included)	**Tender_Price**
Duration	Time (days) to execute the contract	**Duration**
CPV classification	CPV (common procurement vocabulary) is a European system for classifying the type of work in public contracts defined in the Commission Regulation (EC) No 213/2008: http://data.europa.eu/eli/reg/2008/213/oj The numerical code consists of 8 digits, subdivided into divisions (first 2 digits of the code), groups (first 3 digits), classes (first 4 digits), and categories (first 5 digits)	**CPV** **CPV_Aggregated** (first 2 digits of the code)
Contract type	Type of contract defined by legislation (Law 9/2017): works, services, supplies, public works concession, works concession, public services management, services concession, public sector and private sector collaboration, special administrative, private, patrimonial, or others	**Type_code**
Contract subtype	Code to indicate a subtype of contract. If it is a type of service contract: based upon the 2004/18/CE Directive, Annex II. If it is a type of work contract: works contract codes defined by the Spanish DGPE	**Subtype_code**

Table 2: Continued.

Name	Description	Name column dataset
Contract execution place	Contract's execution has a place through the Nomenclature of Statistical Territorial Units (NUTS), created by Eurostat [23]	Not used (assumed equal to Postalzone)
Type of procedure	Procedure by which the contracts was awarded: open, restricted, negotiated with advertising, negotiated without publicity, competitive dialogue, internal rules, derived from framework agreement, project contest, simplified open, association for innovation, derivative of association for innovation, based on a system dynamic acquisition, bidding with negotiation, or others	**Procedure_code**
Contracting system	The contracting system indicates whether it is a contract itself or a framework agreement or dynamic acquisition system	
Type of processing	Type of processing: ordinary, urgent, or emergency	**Urgency_code**
Award result	Type of results: awarded, formalised, desert, resignation, and withdrawal	**Result_code**
Winner identifier	Identifier of the winning bidder (called CIF in Spain) and its province (region)	**CIF_Winner Winner_Province**
Award price	Amount offered by the winning bidder of the contract (taxes included)	**Award_Price**
Date	Date of agreement in the award of the contract	**Date**
Number of received offers	Number of received offers (bidders participating) in each tender	**Received_Offers**

location (latitude and longitude), the municipality, the province, and the autonomous community.

(d) *Conversion*. This consists of converting fields from one format to another. For example, conversions of text fields (strings) to a unique numeric identifier (integers) because the regression algorithm used only works with numeric variables: string_1 => 1, string_2 => 2, ..., string_N = > N.

(5) *Data Filtering*. The data are filtered to discard useless data for our analysis. Basically, this involves the following:

(a) Only formalised or awarded tenders are selected.
(b) A tender is removed when it has one or several empty fields.
(c) A tender is removed when it has an abnormally large positive price (award price or tender price) to remove outliers.
(d) A tender which is formed by several different contracts (called lots) is removed. This is because it does not give the tender price for each contract, and this is a fundamental field for further analysis.

At first, there were 232,175 tenders. After data preprocessing, there were 58,337 tenders.

3. Statistical Analysis of the Dataset

In Section 3.1, a quantitative description of the dataset and a correlation analysis between fields of dataset are presented. In Section 3.2, nine evaluation metrics are defined. In Section 3.3, they are used to calculate the error between two very important fields: tender price versus award price.

3.1. General Description. These data preprocessing operations prepare a structured and organised dataset ready for the data analysis. There are 58,337 tenders from 2012 to 2018 spread across Spain. Table 3 shows the quantitative description of the dataset: total numbers, means, medians, maximum, etc. The dataset has 19 fields or variables: 15 announcement fields and 4 award fields. Special emphasis is placed on Tender_Price and Award_Price. The amount is one of the most important variables in any project. Furthermore, the amount is fundamental in this article because an award price estimator is made.

Looking at Table 3, the following issues are observed:

(i) There are a lot of winning companies and bidding organisations. On average, each public procurement agency makes 16.46 tenders and each company wins 3.37 tenders.

(ii) There is a great dispersion of prices (for both Tender_Price and Award_Price) looking at the median, the mean, and the maximum.

(iii) There is a big difference between Tender_Price and Award_Price looking at the differences between both medians (€14,897) and means (€135,812.48). Therefore, it makes sense to propose a predictor of Award_Price because Tender_Price is not an accurate estimator.

(iv) The 5 types of CPV with greater weight add up to 48.55% of the total number of tenders.

Table 3: Quantitative description of the dataset.

Topic	Description	Value
General values	Total number of tenders in the dataset	58,337
	Temporal range of tenders	2012/01/01–2018/12/28
	Total number of tendering organisations	3,544
	Total number of winning/award companies	17,305
	Mean number of offers received per tender	4.55
	Mean duration of tender's works	382.21 days
Dataset's variables	Input variables of tender's notice: Procedure_code, Urgency_code, Type_code, Subtype_code, Result_code, Name_Organisation, Postalzone, Postalzone_CCAA, Postalzone_Province, Postalzone_Municipality, Tender_Price, CPV, CPV_Aggregated, Duration, and Date	15 input variables (description in Table 2)
	Output variables of tender's resolution: Award_Price, Winner_Province, CIF_Winner, and Received_Offers	4 output variables (description in Table 2)
Tender price (taxes included)	Mean tender price	€538,707.39
	Median tender price	€86,715.00
	Maximum tender price	€3,196,970,000
	Aggregated tender price of all tenders	€31,426,572,936
Award price (taxes included)	Mean award price	€402,894.91
	Median award price	€71,818.00
	Maximum award price	€786,472,000
	Aggregated award price of all tenders	€23,503,680,419
Number of tenders by CPV	Tenders with CPV = 45: construction work	12,166 (20.85%)
	Tenders with CPV = 50: repair and maintenance services	5,174 (8.87%)
	Tenders with CPV = 79: business services (law, marketing, consulting, recruitment, printing, and security)	3,992 (6.84%)
	Tenders with CPV = 72: IT services (consulting, software development, Internet, and support)	3,725 (6.39%)
	Tenders with CPV = 34: transport equipment and auxiliary products to transportation	3,264 (5.60%)
Number of tenders by type code	Tenders with Type_code = 1: goods/supplies	17,876 (30.64%)
	Tenders with Type_code = 2: services	28,363 (48.62%)
	Tenders with Type_code = 3: works	12,008 (20.58%)

To obtain new relevant information through the variables, the Spearman correlation method was used; Figure 2 shows the Spearman correlation matrix (a symmetric matrix with respect to the diagonal). Among the three typical correlation methods (Pearson, Kendall, and Spearman), the Spearman correlation method is chosen because it evaluates the strength of a monotonic relationship between two variables. A monotonic function preserves order (increasing or decreasing). The Spearman correlation coefficient (r_s) is defined for a sample of size n, and the n raw scores X_i, Y_i are converted to ranks rg_{X_i}, rg_{Y_i}:

$$r_s = \frac{\operatorname{cov}(rg_X, rg_Y)}{\sigma_{rg_X}\sigma_{rg_Y}}, \quad (1)$$

where $\operatorname{cov}(rg_X, rg_Y)$ is the covariance of the rank variables and σ_{rg_X} and σ_{rg_Y} are the standard deviations of the rank variables.

Looking at Figure 2, the greatest correlations are the following:

(i) Tender_Price vs. Award_Price (0.97): this high correlation is in accordance with common sense since high bids are associated with high awards and low bids with low awards.

(ii) Type_code vs. Subtype_code (0.74): each type of contract has its associated subtypes of contract. This is the reason for the high correlation.

(iii) Name_Organisation vs. Postalzone_Municipality (0.42): each public procurement agency has a location associated with a postal code.

(iv) Type_code vs. CPV (0.38): each type of contract is usually used for certain types of works.

(v) Procedure_code vs. Tender_Price (−0.38) and Award_Price (−0.36): each type of contract procedure tends to correspond to a range of bidding and adjudication amounts.

(vi) CPV vs. Duration (0.34): each type of work is usually associated with a temporal range (duration) for its realisation.

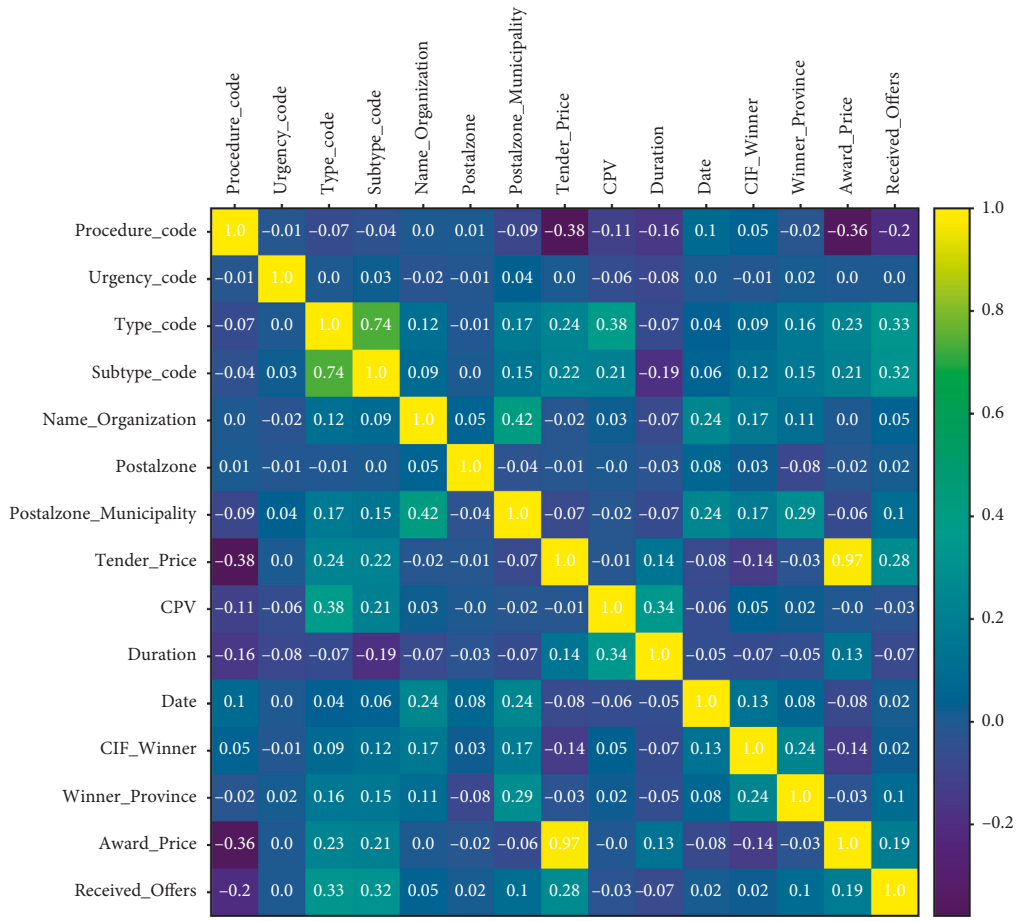


Figure 2: Correlation matrix between the variables of the dataset. Spearman's rank correlation coefficient is the method applied.

(vii) Received_Offers vs. Type_code (0.33) and Subtype_code (0.32): the number of received offers by tender has a correlation with the type and subtype of the contract.

(viii) Winner_Province vs. Postalzone_Municipality (0.29): there is a correlation between the origin (province) of the winning company and the location (municipality) of the tender. In general, tenders from a specific geographical region are won by companies from the same region. There are different socioeconomic reasons for this.

Higher correlation values have not been obtained due to the numerical form of expressing the information and the limitations of the correlation method (all methods have disadvantages). For example, Name_Organisation and Postalzone_Municipality have a direct relation: an organisation usually has a unique assigned postal code. However, this relation can follow any mathematical pattern or function.

Another way to analyse the data is through the scatter matrix (see Figure 3) where the variables are plotted two by two and the matrix's diagonal is the probability density function of the corresponding variable. Although it cannot be appreciated in detail by the large amount of data and variables, the following relations are seen:

(i) Procedure_code, Urgency_code, Type_code, and Subtype_code generate straight lines because they are variables with few values (they are codes) but have great dispersion when they are confronted with the rest of the variables.

(ii) Name_Organisation, Postalzone, and Postalzone_Municipality have a large dispersion. In the probability density function of Postalzone, a great maximum is seen in Madrid's postal codes. This is because many tenders in Spain have been put forward by agencies located in the capital (Madrid).

(iii) The CPVs show that some codes have high tender and award prices, a longer duration, and more received offers. This is true because each type of work has certain characteristics such as price, duration, or competence in the sector.

(iv) The relation between Tender_Price and Award_Price will be analysed in detail later, but a certain relation can be seen. It had already appeared in the correlation matrix.

3.2. Evaluation Metrics. To compare the variables and calculate the errors or deviations of the prediction algorithms, first it is necessary to define some error metrics. The use of

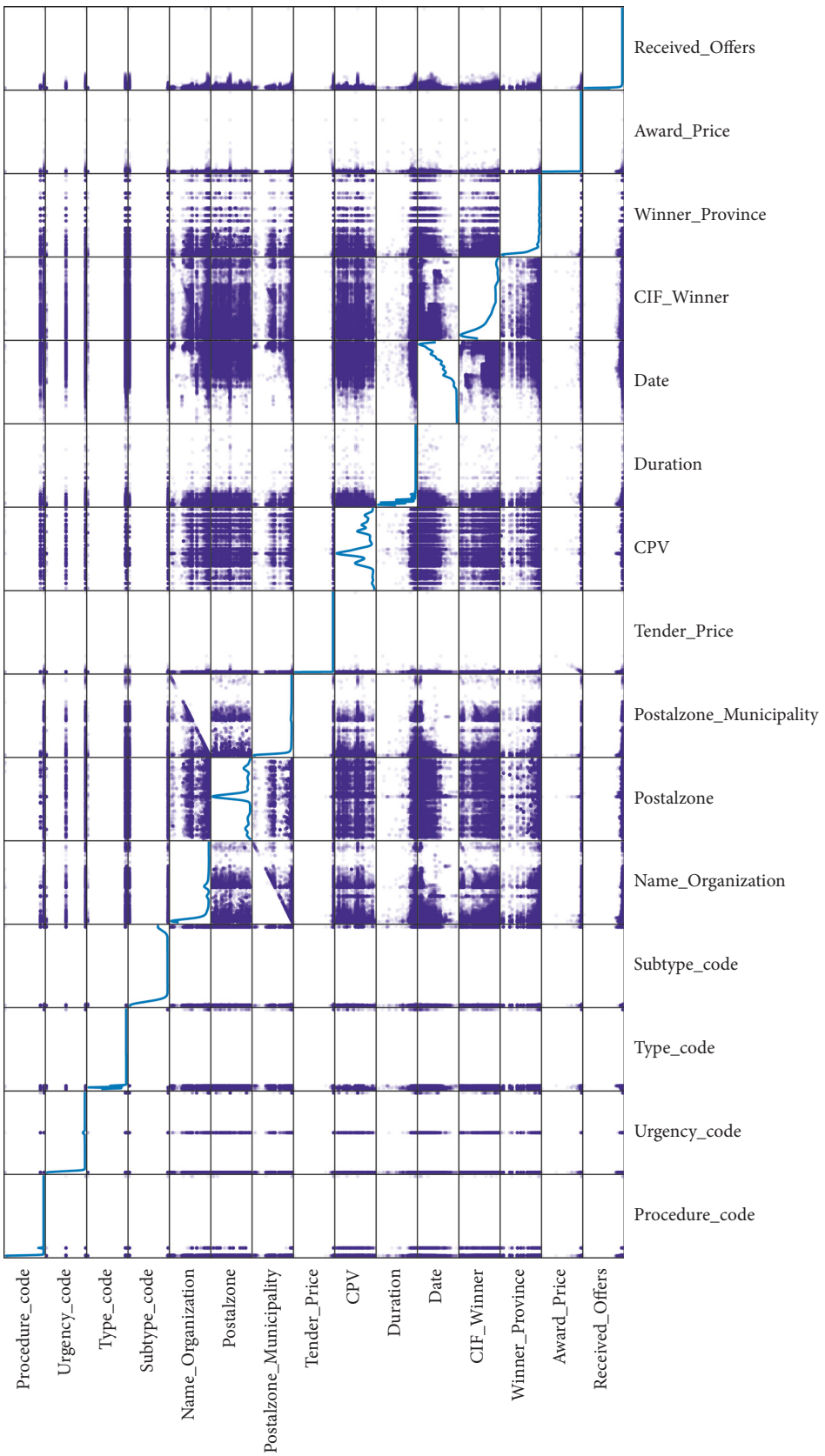
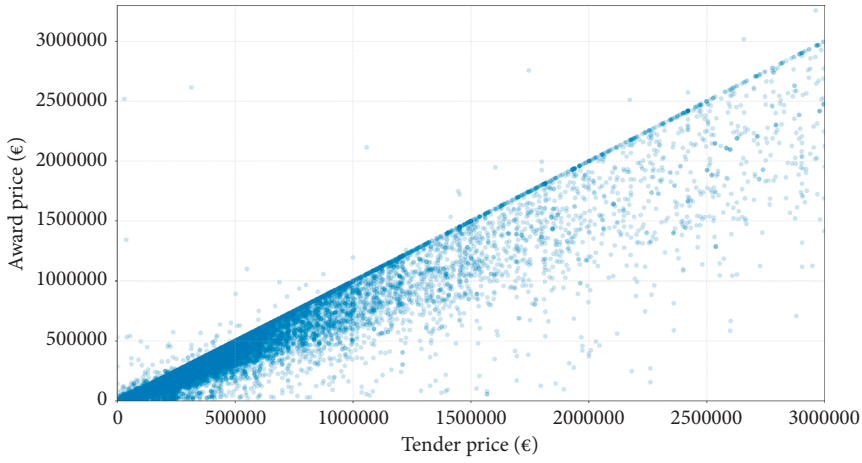
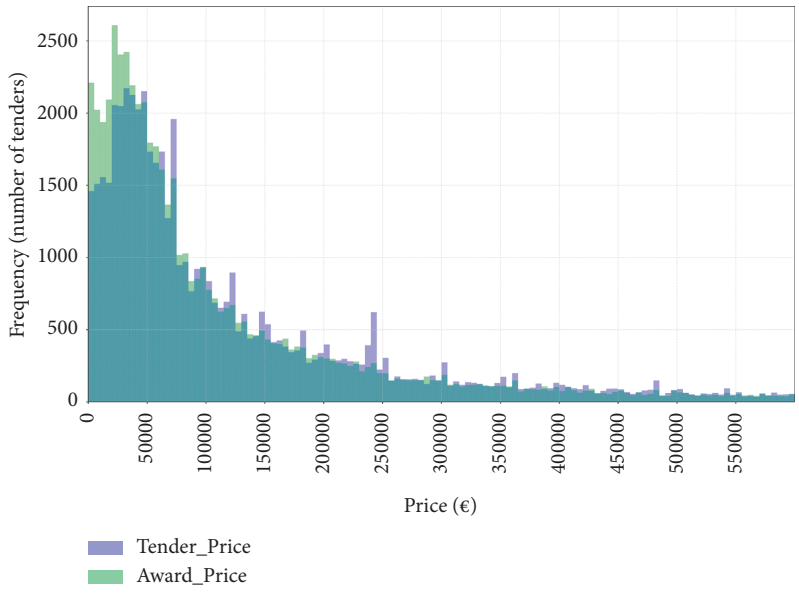


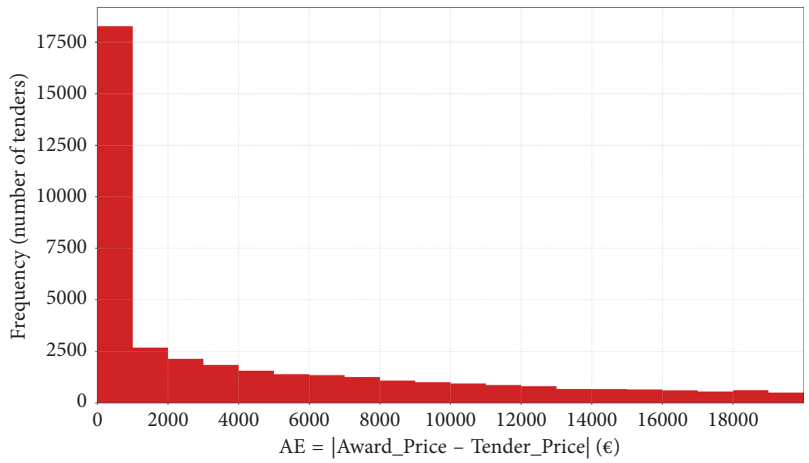
Figure 3: Scatter matrix between the variables of the dataset.



(a)



(b)



(c)

Figure 4: Relation between tender price and award price. (a) Scatter plot. (b) Histograms of frequency (number of tenders). (c) Absolute error (AE) histogram.

metrics based on medians and relative percentage is useful in this survey because the dataset has outliers of great weight, and the use of such metrics helps us to counteract the effect of these outliers.

Absolute error (AE), absolute percentage error (APE), mean absolute error (MAE), mean absolute percentage error (MAPE), median absolute error (MdAE), median absolute percentage error (MdAPE), root mean square error (RMSE), normalised root mean square error (NRMSE), and coefficient of determination (R^2) were selected as evaluation criteria (2)–(10): A_t is the actual value for period t, F_t is the expected or estimated value for period t, and n is the number of periods.

$$\mathrm{AE}_t = \left|A_t - F_t\right|, \tag{2}$$

$$\mathrm{APE}_t\ (\%) = 100\left|\frac{\mathrm{AE}_t}{A_t}\right| = 100\left|\frac{A_t - F_t}{A_t}\right|, \tag{3}$$

$$\mathrm{MAE} = \frac{1}{n}\sum_{t=1}^{n}\mathrm{AE}_t = \frac{1}{n}\sum_{t=1}^{n}\left|A_t - F_t\right|, \tag{4}$$

$$\mathrm{MAPE}\ (\%) = \frac{100}{n}\sum_{t=1}^{n}\mathrm{APE}_t = \frac{100}{n}\sum_{t=1}^{n}\left|\frac{A_t - F_t}{A_t}\right|, \tag{5}$$

$$\mathrm{MdAE} = \frac{1}{n}\mathrm{median}\left(\left|A_1 - F_1\right|, \left|A_2 - F_2\right|, \ldots, \left|A_n - F_n\right|\right), \tag{6}$$

$$\mathrm{MdAPE}\ (\%) = \frac{100}{n}\mathrm{median} \cdot \left(\left|\frac{A_1 - F_1}{A_1}\right|, \left|\frac{A_2 - F_2}{A_2}\right|, \ldots, \left|\frac{A_n - F_n}{A_n}\right|\right), \tag{7}$$

$$\mathrm{RMSE} = \sqrt{\frac{1}{n}\sum_{t=1}^{n}\left|A_t - F_t\right|^2}, \tag{8}$$

$$\mathrm{NRMSE} = \frac{\mathrm{RMSE}}{\max\left(A_t\right) - \min\left(A_t\right)} = \frac{\sqrt{(1/n)\sum_{t=1}^{n}\left|A_t - F_t\right|^2}}{\max\left(A_t\right) - \min\left(A_t\right)}, \tag{9}$$

$$R^2 = 1 - \frac{\sum_{t=1}^{n}\left|A_t - F_t\right|^2}{\sum_{t=1}^{n}\left|A_t - \overline{A}\right|^2}, \tag{10}$$

where $\overline{A}$ is the mean: $\overline{A} = (1/n)\sum_{t=1}^{n}A_t$.

3.3. Tender Price vs. Award Price. Figure 4(a) shows graphically the variable tender price versus award price for all tenders when tender price is less than €3,000,000. This threshold is 3.5 times the median of tender price. A line at 45 degrees can be seen; its points satisfy the condition that tender price is equal to award price. Therefore, in this line there, is no error between the two variables, and so the tender price would be a perfect estimator. Below this line, there is a large dispersion of points. When the distance between a point and the line is high, the error is also high. Finally, there are few points above the line. This is because only rarely is the award price higher than the tender price. This can happen due to special conditions of the contract or, alternatively, it can be wrong data. There is no information about how the public procurement agencies calculate the tender price or if it is validated before entering the dataset.

Table 4: Error metrics between tender price and award price.

Error	Value
Absolute error (AE)	See Figure 4(c)
Absolute percentage error (APE)	See Figure 5
Median absolute error (MdAE)	€6,955.00
Median absolute percentage error (MdAPE)	11.84%
Mean absolute error (MAE)	€137,778.64
Mean absolute percentage error (MAPE)	39.79%
Root mean square error (RMSE)	101,451,609,620,714
Coefficient of determination	−3.10

Figure 4(b) shows the frequency histogram of both variables. The frequency is the number of tenders for each bar of €5,000. For example, the most frequent range for the tender price is €30,000–€35,000; for the award price, it is €20,000–€25,000. Figure 4(c) shows the frequency histogram of the AE between both variables by ranges of €1,000. It can be observed that approximately 18,000 tenders (30% of the total) have less than €1,000 error. There is a big difference with the rest of the bars.

Table 4 presents the error metrics (or evaluation metrics) calculated between the variables tender price and award price for the entire dataset. An error between tender price and award price, in terms of project management, means that there is a budget deviation between the tender price and the price finally awarded.

An interesting analysis is how the award price is affected by the competitiveness of the companies (see Table 5). It is necessary to group the tenders according to the number of offers received. For this purpose, 4 groups have been created: no competitiveness (1 offer), low competitiveness (2–4 offers), medium competitiveness (5–10 offers), and high competitiveness (more than 10 offers). As competitiveness increases, the difference between the award price and tender price is greater because MdAE, MdAPE, MAE, and MAPE are greater. This shows that companies are more aggressive (bid lower prices) to win the tender. Consequently, the award price is lower in a scenario with less competitiveness or, in other words, public procurement agencies save money.

Figure 5 shows the APE boxplot grouped by CPV. Box diagrams are a standard method to graphically represent numerical data through their quartiles. The outliers of the dataset have not been represented because they are values very far out, which would make it difficult to scale the axes. MAPE (red colour) and MdAPE (green colour) for each CPV group are marked. The great differences of APE,

Table 5: Description of the dataset and the errors between tender price and award price by number of received offers.

Description	Groups by competitiveness			
	No competitiveness	Low Received offers (2–4)	Medium Received offers (5–10)	High Received offers >10
Total number of tenders in the dataset	18,790	22,714	11,553	5,271
Total number of tendering organisations	1,956	2,553	2,135	1,053
Total number of winning/award companies	7,550	9,555	5,222	2,402
Mean received offers by tender	1.0	2.80	6.73	20.01
Mean duration of tender's works	401.07 days	396.65 days	370.95 days	277.50 days
Mean tender price	€354,882.49	€388,526.27	€785,455.49	€1,301,031.70
Median tender price	€60,500.00	€75,000.00	€121,000.00	€254,376.00
Mean award price	€341,874.79	€323,611.87	€460,548.68	€836,188.79
Median award price	€58,984.50	€64,833.00	€90,689.00	€174,986.00
Median absolute error (MdAE)	€93.50	€7,661.50	€22,854.00	€76,420.00
Median absolute percentage error (MdAPE)	0.12%	13.39%	29.63%	45.94%
Mean absolute error (MAE)	€13,966.65	€68,244.60	€326,698.33	€464,907.75
Mean absolute percentage error (MAPE)	10.02%	25.65%	54.48%	77.98%

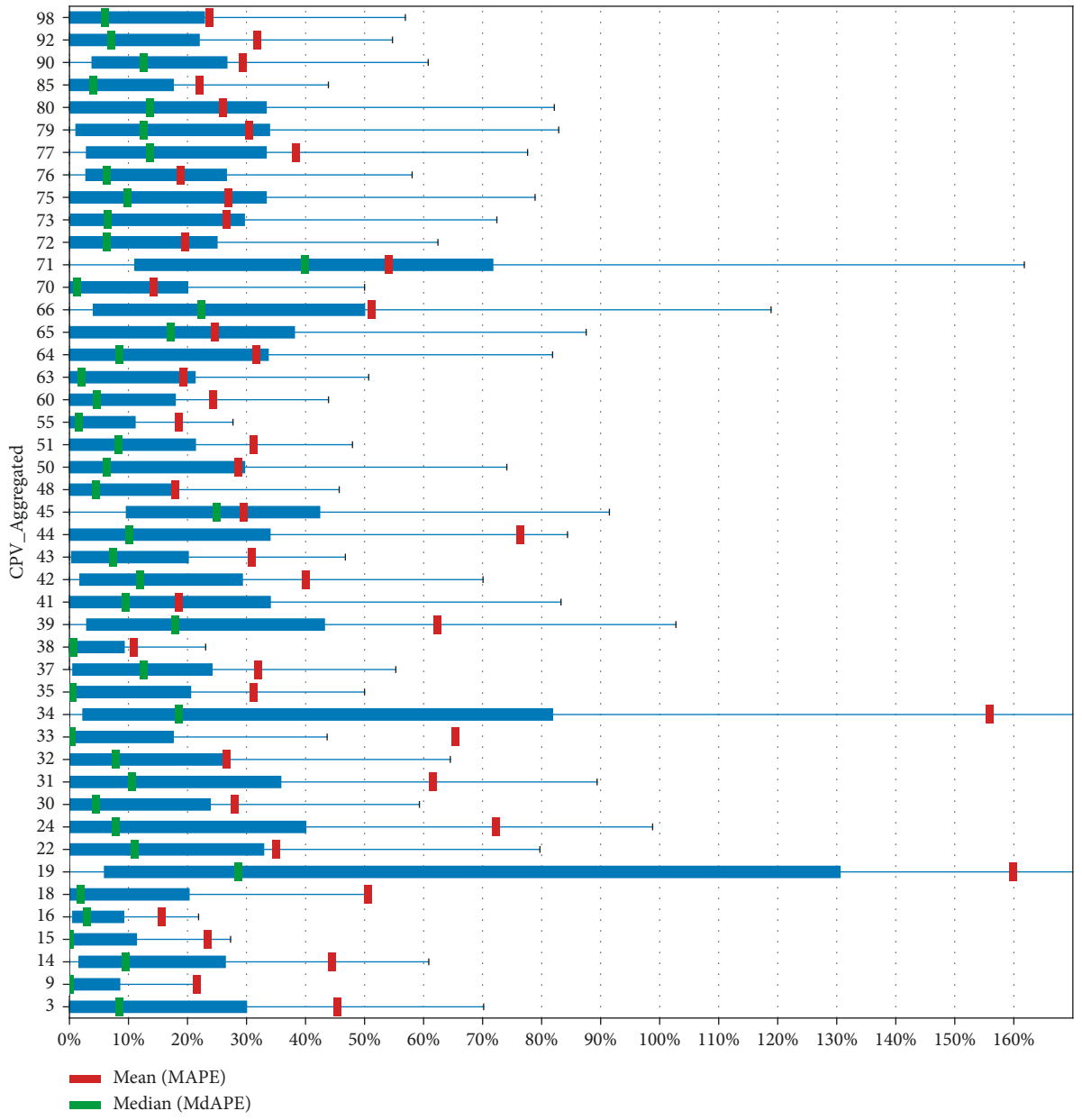


Figure 5: Boxplot of absolute percentage error (APE) between award price and tender price grouped by CPV.

MAPE, and MdAPE according to the CPV can be clearly seen. In general, MdAPE is between 20% and 40% and MAPE is higher than 40%. The total value of MAPE and MdAPE (without dividing by CPV) has already been calculated, as shown in Table 4.

In conclusion, in view of the graphical and quantitative results, it can be affirmed that tender price is a bad estimator of award price. Perhaps it is not excessively bad in median (11.84%) but it is so in mean (39.79%). This is certainly due to the high dispersion between both prices (as seen in Figure 4(a)). This is the reason to create an award price estimator in the following section.

4. Award Price Estimator

A good award price estimator would be very useful and valuable for companies and public procurement agencies. It would be a key tool in their project management decision making because it reduces the economic risks. Due to the complexity involved, machine learning techniques have been chosen to create the estimator, in particular, random forest. In Section 4.1, random forest for regression is presented, from the theoretical framework to its application to the Spanish tenders' dataset. In Section 4.2, the empirical results and analysis are presented, for example, the error metrics of the award price estimator created. In Section 4.3 a similar analysis is presented using a dataset from other countries, creating a new award price estimator.

4.1. Random Forest for Regression. Random forests (RF), introduced by Breiman [24] in 2001, is an ensemble learning method for regression or classification that operates by constructing a multitude of decision trees at training time and outputting the class which is the mode of the classes (classification) or mean prediction (regression) of the individual trees. It is a popular learning algorithm that offers excellent performance [25, 26], no overfitting [27], and a versatility of applicability to large-scale problems and in handling different types of data [25, 28]. It provides its own internal generalisation error estimate, called out-of-bag (OOB) error.

Simplified algorithm of RF for regression [29]:

(1) For $b = 1$ to B:

  (a) Draw a bootstrap sample $\mathbb{Z}^*$ of size N from the training data.
  (b) Grow a random forest tree T_b to the bootstrapped data, by recursively repeating the following steps for each terminal node of the tree, until the minimum node size $n_{\min}$ is reached.
    (i) Select m variables at random from the p variables.
    (ii) Pick the best variable/split-point among the m.
    (iii) Split the node into two daughter nodes.

(2) Output the ensemble of trees $\{T_b\}_1^B$.

To make a prediction at a new point x, $\widehat{f}_{rf}^{B}(x) = (1/B)\sum_{b=1}^{B} T_b(x)$.

At each split in each tree, the improvement in the split criterion is the measure of the importance attributed to the splitting variable and is accumulated over all the trees in the forest separately for each variable. It is called variable importance [24].

There are other implementations of RF algorithms, such as Boruta [30], regularised random forest (RRF) [31], conditional forest [32], quantile regression forest (QRF) [33], or extremely randomised regression trees (extraTrees) [34]. The last one was tested with this dataset, but it has a worse accuracy than random forest, so finally it was discarded. The reason is because the function to measure the quality of a split is the Gini index, which is worse than MAE (mean absolute error) or MSE (mean squared error). A comparison between the use of MAE and MSE is shown in Figure 6 for 30 to 1000 trees generated in RF. MAE used as the quality function has clearly better values for the error metrics (especially MAPE and NRMSE) than the MSE quality function for this dataset. Therefore, the function selected is MAE.

The random forest method has been used for multiple and different real-world applications [25], such as the estimation of traffic car issues [35–37], wind speed prediction [38], classification of protein sequences [39], discrimination between seismic events and nuclear explosions [40], pedestrian detection [41], aggregated recommender systems [42], bed occupancy predictor in hospitals [43], classification of phishing e-mail [44], network intrusion detection [45], and employee turnover prediction [46].

Figure 7 shows different ratios between the training and testing subsets (train : test in percentage): 65 : 35, 70 : 30, 75 : 25, 80 : 20, 85 : 15, and 90 : 10. The most important errors for this study, MdAPE and MAPE, are constantly in the order of 9% and 30%, respectively. OBB and NMRSE do not change significantly. Hence, the train : test ratio is not relevant. The typical ratio 80 : 20 will be used in this article.

RandomForestRegressor from *Scikit-learn*, which is a machine learning library for the Python programming language, with 400 trees is the function used in this article. The 14 input variables used in RF are Tender_Price, Date, Duration, Name_Organisation, CPV, CPV_Aggregated, Procedure_code, Type_code, Subtype_code, Urgency_code, Postalzone, Postalzone_CCAA, Postalzone_Province, and Postalzone_Municipality. The variable to perform the regression is Award_Price, and the output generated by RF (prediction) will be called Forecast_Price.

This article does not use the other 3 variables of the tender's resolution (Winner_Province, CIF_Winner, and Received_Offers; Table 3) because they are not variables of the tender's notice. In a real scenario, the award price estimator only can use the variables of the tender's notice. However, if these 3 output variables are used in RF plus 14 input variables, the errors would decrease logically. This is demonstrated as shown in Figure 8: MdAPE is about 5% and MAPE 25%. MdAPE and MAPE are, respectively, 4% and 5% lower than the real scenario with only variables of the tender's notice (see Figure 7). The variable importances (RF

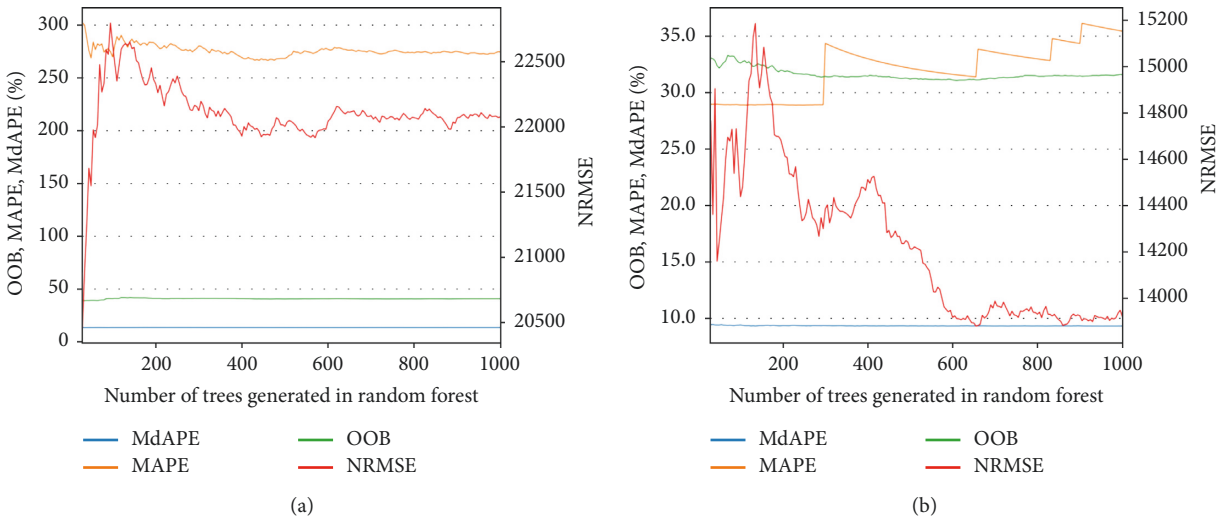


Figure 6: Relationship between trees in random forests (number of estimators) and error metrics (MdAPE, MAPE, OBB, and NRMSE) for two functions to measure the quality of a split. (a) The quality function is mean squared error (MSE). (b) The quality function is mean absolute error (MAE).

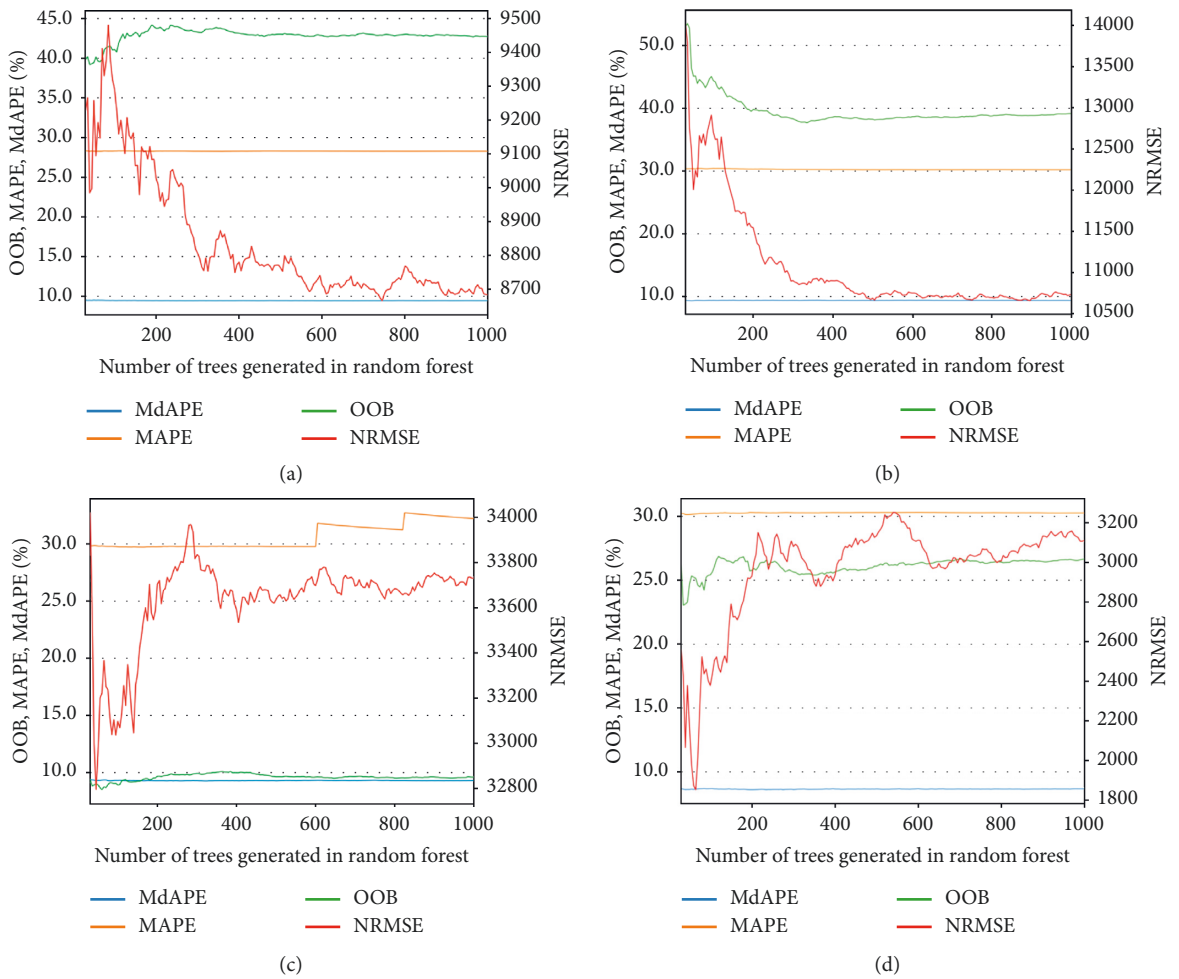


Figure 7: Continued.

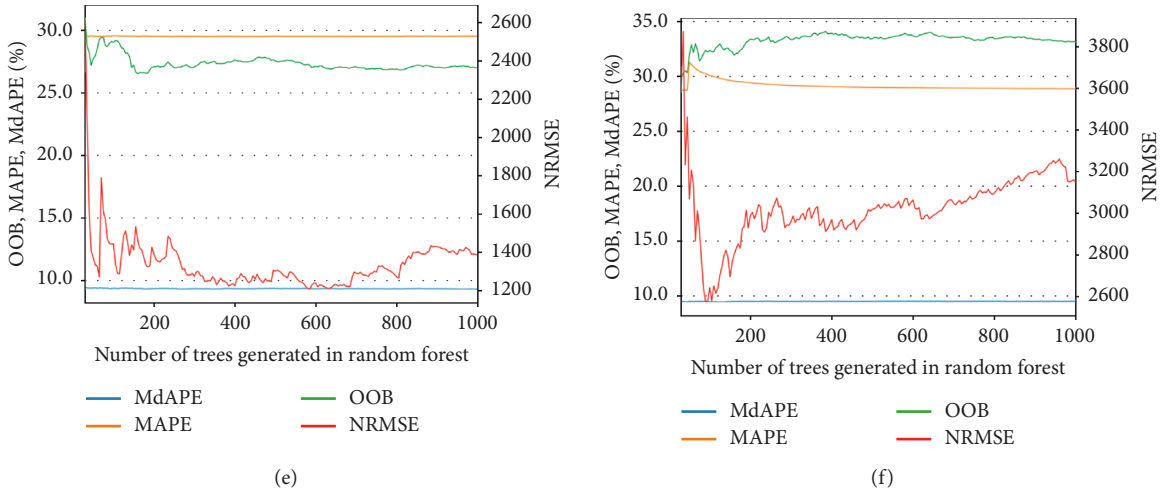


Figure 7: Relationship between trees in random forests and error metrics (MdAPE, MAPE, OBB, and NRMSE) for different ratios of training and testing subsets. (a) 65 : 35. (b) 70 : 30. (c) 75 : 25. (d) 80 : 20. (e) 85 : 15. (f) 90 : 10.

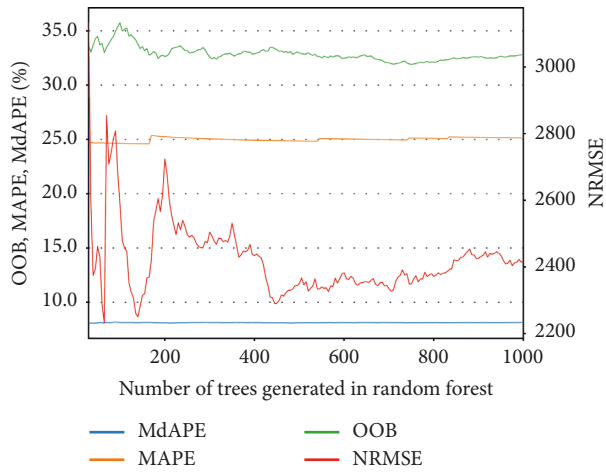


Figure 8: Relationship between trees in random forests and error metrics (MdAPE, MAPE, OBB, and NRMSE) using the 14 input variables plus 3 variables of the tender's resolution.

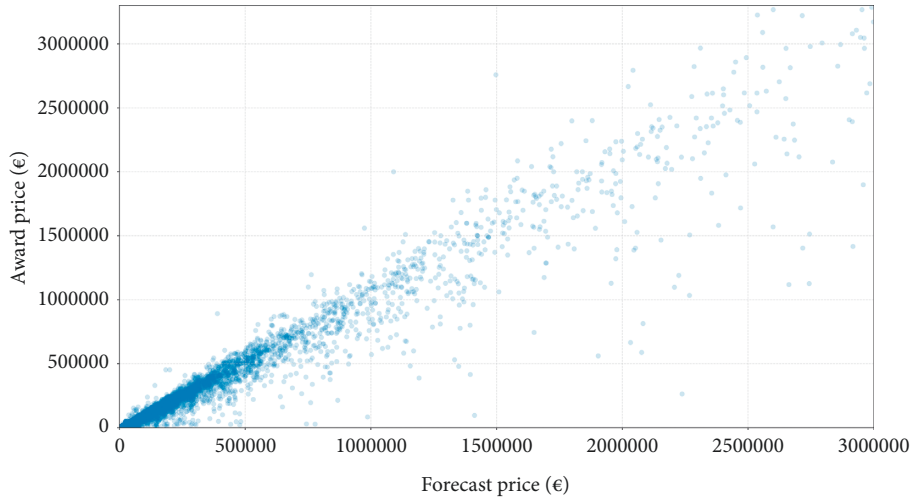


Figure 9: Scatter plot between forecast price and award price.

Table 6: Error metrics between award price and forecast price.

Error	Value	Difference with respect to Tender_Price
Absolute percentage error (APE)	See Figure 8	See Figure 8
Median absolute error (MdAE)	€7,575.45	+€620.45
Median absolute percentage error (MdAPE)	9.26%	−2.58%
Mean absolute error (MAE)	€67,241.34	+€70,537.3
Mean absolute percentage error (MAPE)	28.60%	−11.19%
Root mean square error (RMSE)	364,901,707,583	−101,086,707,913,131
Coefficient of determination (R^2)	0.92	—

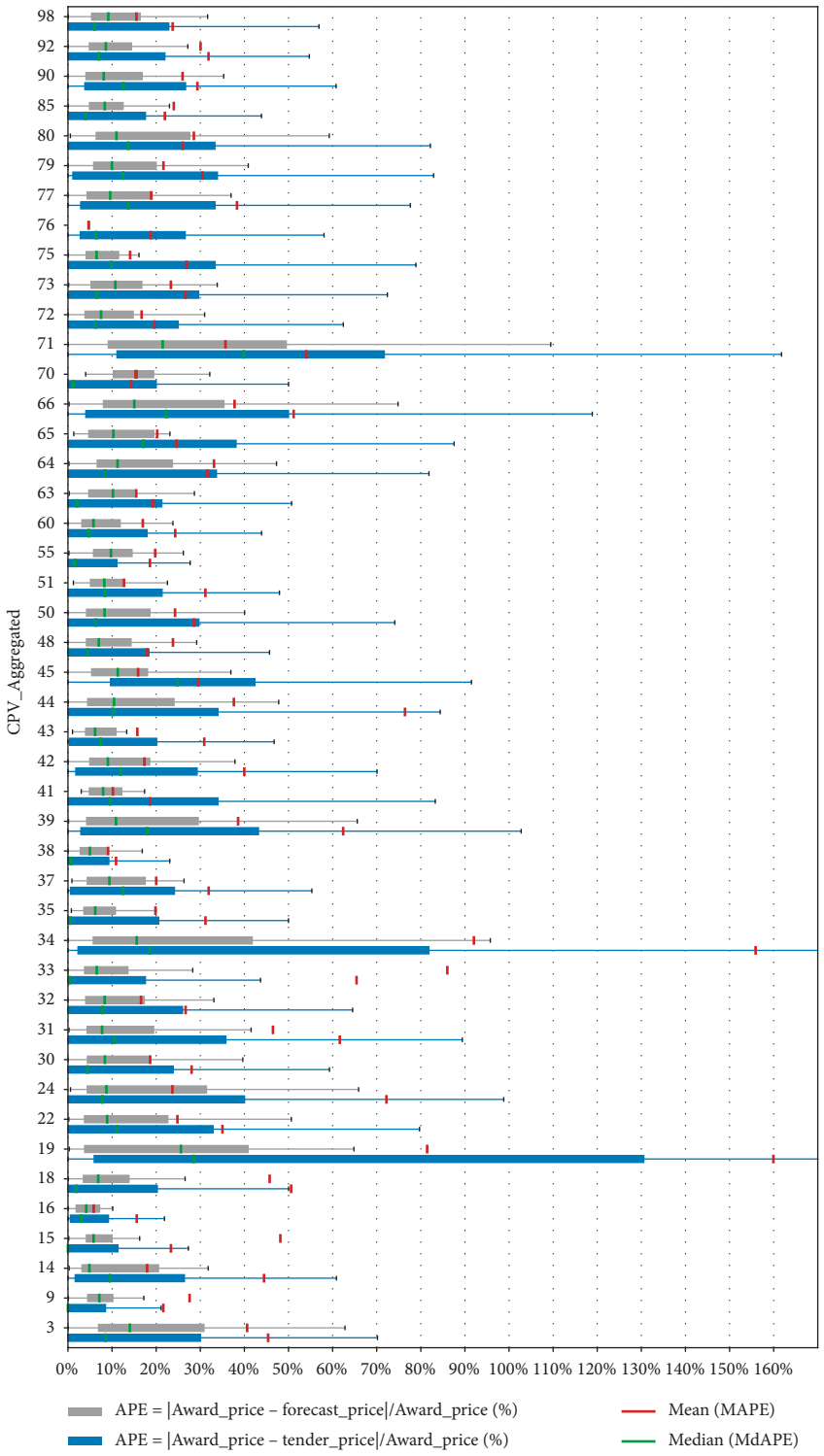


Figure 10: Boxplot of absolute percentage error (APE, grey colour) between award price and forecast price, grouped by CPV. The APE reference (blue colour) is the award price and tender price shown in Figure 5.

Table 7: European countries' dataset: quantitative description.

Topic	Description	Value
General values	Total number of tenders in the dataset	41,556
	Number of tenders by country: France (FR), Croatia (HR), Slovenia (SI), Bulgaria (BG), Germany (DE), Italy (IT), Hungary (HU), and Latvia (LV)	12,449 (FR); 7,910 (HR); 6,473 (SI); 6,096 (BG); 3,918 (DE); 3,782 (IT); 3,724 (HU); 1,736 (LV)
	Temporal range of tenders	2016/12/22–2017/12/29
	Total number of tendering organisations	6,163
	Total number of winning/award companies	19,100
	Mean received offers by tender	5.02
Dataset's variables	Input variables of tender's notice: Date, Name_Organisation, Postalzone, ISO_country_code, Main_activity, Type_code, CPV, CPV_Aggregated, Tender_Price, and Procedure_code	10 input variables
	Output variables of tender's resolution: Award_Price	1 output variable
Prices (without taxes)	Median tender price	€425,000.00
	Median award price	€394,951.26
Number of tenders by CPV	Tenders with CPV = 33: medical equipments, pharmaceuticals, and personal care products	10,927 (26.29%)
	Tenders with CPV = 15: food, beverages, tobacco, and related products	4,363 (10.50%)
	Tenders with CPV = 45: construction work	4,053 (9.75%)
	Tenders with CPV = 71: architectural, construction, engineering, and inspection services	1,973 (4.75%)
	Tenders with CPV = 34: transport equipment and auxiliary products to transportation	1,893 (4.56%)
Number of tenders by type code	Tenders with Type_code = 1: goods/supplies	24,593 (59.18%)
	Tenders with Type_code = 2: services	12,849 (30.92%)
	Tenders with Type_code = 3: works	4,114 (9.90%)

Table 8: European countries' dataset: errors between award price vs. tender price and award price vs forecast price and their differences.

Error	Award price vs. tender price	Award price vs. forecast price	Difference
Median absolute error (MdAE)	€4,514.50	€20,982.94	+€16,468.44
Median absolute percentage error (MdAPE)	4.17%	6.48%	+2.31%
Mean absolute percentage error (MAPE)	27.49%	23.57%	−3.92%
Normalised root mean square error (NRMSE)	99,018.04	2,816,245.06	+2,717,227.02
Coefficient of determination (R^2)	0.9680	0.7303	−0.2377

output parameter) ordered from highest to lowest are Tender_Price (0.870%), Received_Offers (0.035%), Duration (0.017%), Date (0.013%), Name_Organisation (0.012%), CIF_Winner (0.010%), CPV (0.009%), Postalzone (0.007%), Subtype_code (0.006%), CPV_Aggregated (0.005%), Winner_Province (0.004%), Type_code (0.004%), Procedure_code (0.003%), Postalzone_Municipality (0.002%), Postalzone_Province (0.001%), Postalzone_CCAA (0.001%), and Urgency_code (0.0001%). It is clear that the 3 output variables are important in the previous ranking.

4.2. Empirical Results and Analysis. RF has been trained with 80% of tenders (46,670). The remaining 20% (11,667) have been used as the test group. Figure 9 shows the scatter plot between forecast price and award price for the test group. As has already been mentioned, if the estimator were perfect, all points would have to be on the line at 45 degrees.

The prediction's errors are presented in Table 6. Furthermore, in the third column, it is compared with the error made by Tender_Price (see Table 4) to check if the proposed estimator is better or worse. It makes no sense to compare the absolute errors because the sizes of the datasets are different. It is best to compare the percentage errors, such as MdAPE and MAPE; they are significantly lower, MdAPE—2.58% and MAPE—11.19%.

Figure 10 shows the boxplot of APE (grey colour) between award price and forecast price grouped by CPV. It is also plotted the APE reference (blue colour) which has been presented previously in Figure 5. It is clearly visible how the APE of the estimator has boxplots with a smaller interquartile range (IQR). In general, MdAPE and MAPE are lower than the APE reference. In conclusion, the proposed estimator reduces significantly the error with respect to tender price (analysed in Section 3.3).

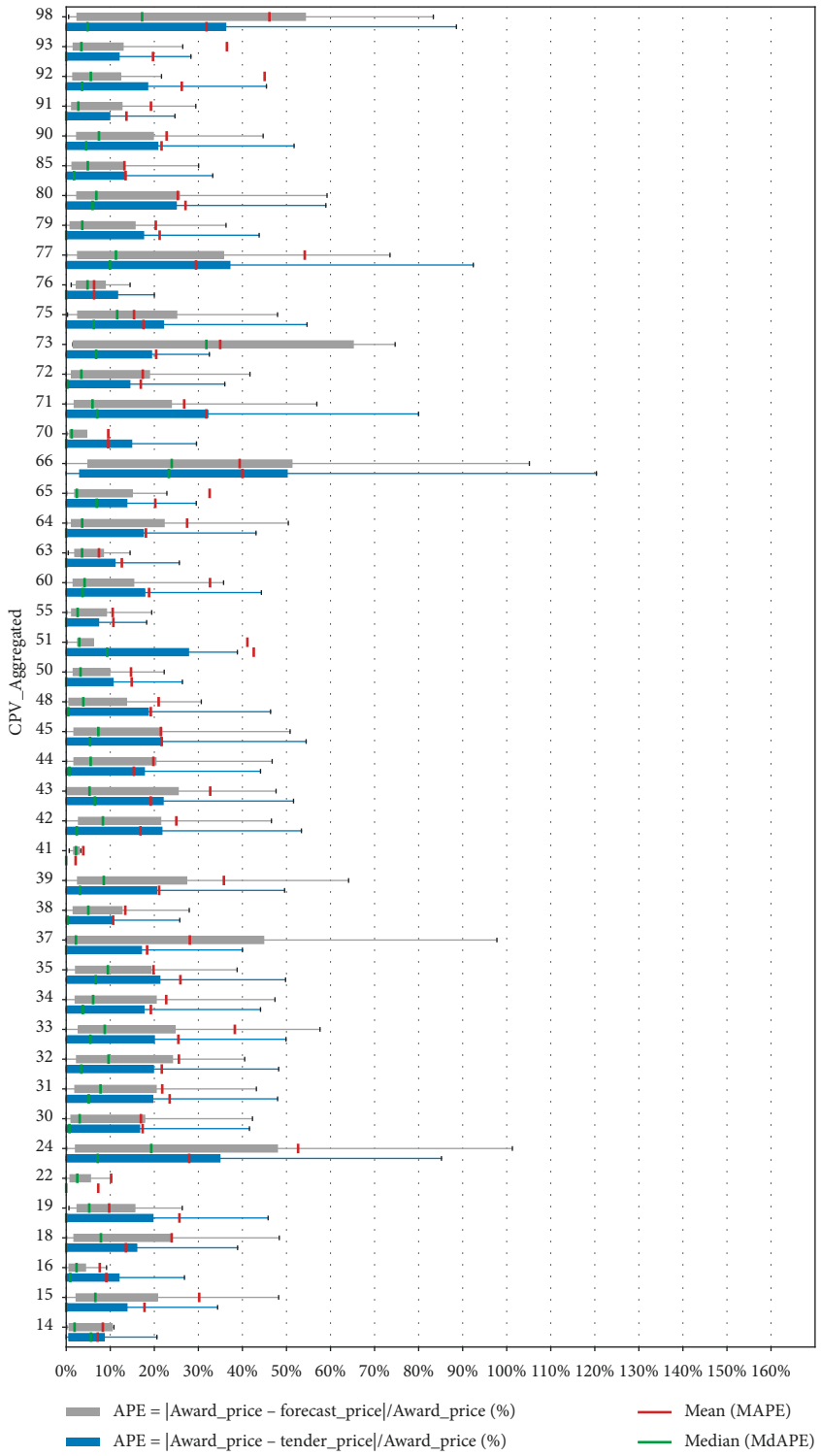


Figure 11: European countries' boxplot: absolute percentage error (APE) between award price and tender price (blue colour) and award price and forecast price (grey colour), grouped by CPV.

The variable importances (RF output parameter) ordered from highest to lowest are Tender_Price (88.34%), Date (1.84%), Duration (1.83%), Name_Organisation (1.56%), Subtype_code (1.52%), CPV (1.10%), Postalzone (1.09%), Type_code (0.97%), Procedure_code (0.66%), CPV_Aggregated (0.49%), Postalzone_Municipality (0.24%), Postalzone_Province (0.18%), Postalzone_CCAA (0.17%), and Urgency_code (0.03%).

4.3. Empirical Results and Analysis for Other Countries. In this section, a study is made with tenders from other countries, similar to the previous one for Spanish tenders. The

purpose is to evaluate the award price estimator with a different dataset using the same machine learning technique (random forest). The countries selected are from the European Union because they have almost the same characteristics associated with public procurement announcements: legislation, tender's regulation, public administrations, purchase procedures, etc. The raw data have been downloaded from the European Open Data Portal [21], in particular the tenders' database of the year 2017 (link in the Data Availability section). However, the quality of the data is not good: fields without data, errors in tender and award prices, the winning company does not have its tax identification number, tender and award prices have the same value, etc. It is an official dataset provided by the European Union, but it does not have as good a quality as the Spanish dataset. In the beginning, there were 706,104 tenders. After data preprocessing, there were only 41,556 tenders.

Table 7 shows the quantitative description of the dataset for the following 8 European countries: France, Germany, Italy, Croatia, Slovenia, Bulgaria, Hungary, and Latvia. They have been selected because they have the highest number of tenders after data preprocessing.

This dataset has been trained with 80% of the tenders (33,244). The remaining 20% (8,312) have been used as the test group. The random forest process is analogous to the Spanish one. The 10 input variables used in RF are Date, Name_Organisation, Postalzone, ISO_country_code, Main_activity, Type_code, CPV, CPV_Aggregated, Tender_Price, and Procedure_code. The variable to perform the regression is Award_Price, and the output generated by RF (prediction) will be called Forecast_Price.

The errors MdAE, MdAPE, MAPE, and NRMSE and R^2 are shown in Table 8. The second column shows award price vs. tender price (the reference), and the third column shows award price vs. forecast price (the estimator created with RF). MdAPE between award price and tender price is very low (4.17%) if it is compared to the Spanish MdAPE (11.84%, see Table 4). This means that award price is very close to tender price or, in other words, a lot of tenders have the same price for both and, consequently, without error. MAPE is also lower (27.49%) than the Spanish MAPE (39.79%). The estimator is better in MAPE (−3.92%) but it is worse in MdAPE (+2.31%) (see fourth column in Table 8).

Figure 11 shows the boxplot of APE (grey colour) between award price and forecast price grouped by CPV. The APE reference (blue colour) between award price and tender price is also plotted. It is not clearly visible how the APE of the estimator has boxplots with a smaller interquartile range (IQR). In general, MdAPE and MAPE are similar to the APE reference.

In conclusion, the estimator created for this dataset has similar error metrics with respect to tender price. Why do a lot of tender notices in the European countries have the same value of tender price and award price? Why not in the Spanish case? This could be due to the bad quality of the European dataset (tender's notices with mistakes) or, a less likely hypothesis, the fact that the Spanish public procurement agencies fail to estimate the tender price and the European agencies never fail in anything. The proposed method can be useful and generalisable to other countries with a large dataset without mistakes.

5. Conclusions and Future Research

The European and Spanish public procurement legislation has been presented. A dataset of 58,337 Spanish public tenders from 2012 to 2018 has been analysed. The relations between the main fields of the public procurement notices have been studied mathematically. Error metrics between the tender price and the award price have been calculated (MdAPE = 11.84% and MAPE = 39.79%). An award price estimator, which reduces the previous errors (MdAPE = 9.26% and MAPE = 28.60%), has been proposed by using a machine learning algorithm (random forest). The estimator has 14 fields as input variables, of which the most important are the tender price, date, duration, public procurement agency name, subtype code, CPV classification, and postal zone.

A good award price estimator would be useful for companies and public procurement agencies. It would be useful for companies because it can be a key tool in their project management decision making: it would reduce the economic risks, thus winning tenders more easily. For public procurement agencies, it would be useful because, for example, in the Spanish dataset, the tender price could have been reduced by 2.24% (MdAPE reduction), equivalent to approximately 811 million euros. This is a significant error reduction that, consequently, would improve the accuracy of the budget for public procurement.

An analogous analysis has been made with 8 European countries (France, Germany, Italy, Croatia, Slovenia, Bulgaria, Hungary, and Latvia) to generalise the award price estimator to other real situations and check the results. The dataset used has 41,556 tenders, but the quality of the data is worse than the Spanish dataset. The new award price estimator obtains predictions with error metrics that are similar to those between the tender price and award price. The estimator is better in MAPE (−3.92%) but it is worse in MdAPE (+2.31%).

An accurate estimate is impossible to achieve because the market is theoretically open and free and, therefore, unpredictable. Furthermore, the award price is not always the final price paid by the public procurement agency because the contract may be modified during its execution. However, this article illustrates how a machine learning algorithm can be useful. Particularly, random forest predicts the award prices with less uncertainty, adapting to the real market. This market reality is gathered implicitly through the public procurement notices. Therefore, this estimator is interesting for the public procurement agencies and for the companies because their risk is reduced.

Thanks to the open data sources of public procurement, it is possible to avoid depending on government statistics offices such as the Spanish (INE [47]) or the European (Eurostat [23]). Therefore, there is independence, and there are resources to perform low-level analysis or cross data with

other databases or external services to extract more valuable information.

This article opens the doors to future research related to the analysis of massive data on public procurement, in particular:

(i) It achieves a more accurate estimator by incorporating business data of the winning bidder: location, core business, annual turnover, number of employees, financial situation, etc. With the new data, the estimator has more input variables that could be relevant to predicting the award price.

(ii) It compares other machine learning algorithms to estimate award prices, number of received offers, and other interesting fields.

(iii) It performs data business analysis such as companies with a higher success rate in public procurement or the characterisation of the winning bidder: type of company, size, national origin or foreign, etc.

Data Availability

The processed data used to support the findings of this study are available from the corresponding author upon request. The raw data from Spain are available at the Ministry of Finance, Spain. Open data of Spanish tenders are hosted in http://www.hacienda.gob.es/es-ES/GobiernoAbierto/Datos%20Abiertos/Paginas/licitaciones_plataforma_contratacion.aspx. The raw data from other countries are available in the European Union Open Data Portal (Publications Office of the European Union) hosted in https://data.europa.eu/euodp/en/data/dataset/ted-csv.

Conflicts of Interest

The authors declare that there are no conflicts of interest regarding the publication of this paper.

Acknowledgments

This work was supported by the Plan of Science, Technology and Innovation of the Principality of Asturias (Ref: FC-GRUPIN-IDI/2018/000225).

References

[1] European Commission, "European semester thematic factsheet public procurement," 2017, https://ec.europa.eu/info/sites/info/files/file_import/european-semester_thematic-factsheet_public-procurement_en_0.pdf.

[2] The National Securities Market Commission (CNMV) from Spain, "Radiography of public procurement procedures in Spain," 2019, https://www.cnmc.es/sites/default/files/2314114_5.pdf.

[3] N. Obwegeser and S. D. Müller, "Innovation and public procurement: terminology, concepts, and applications," *Technovation*, vol. 74-75, pp. 1–17, 2018.

[4] J. Attard, F. Orlandi, S. Scerri, and S. Auer, "A systematic review of open government data initiatives," *Government Information Quarterly*, vol. 32, no. 4, pp. 399–418, 2015.

[5] H. R. Varian, "Big data: new tricks for econometrics," *Journal of Economic Perspectives*, vol. 28, no. 2, pp. 3–28, 2014.

[6] S. Mullainathan and J. Spiess, "Machine learning: an applied econometric approach," *Journal of Economic Perspectives*, vol. 31, no. 2, pp. 87–106, 2017.

[7] J. M. Alvarez-Rodríguez, J. E. Labra-Gayo, and P. O. De Pablos, "New trends on e-procurement applying semantic technologies: current status and future challenges," *Computers in Industry*, vol. 65, no. 5, pp. 800–820, 2014.

[8] M. Nečaský, J. Klímek, J. Mynarz, T. Knap, V. Svátek, and J. Stárka, "Linked data support for filing public contracts," *Computers in Industry*, vol. 65, no. 5, pp. 862–877, 2014.

[9] J. D. Twizeyimana and A. Andersson, "The public value of e-government—a literature review," *Government Information Quarterly*, vol. 36, no. 2, pp. 167–178, 2019.

[10] M. A. Bergman and S. Lundberg, "Tender evaluation and supplier selection methods in public procurement," *Journal of Purchasing and Supply Management*, vol. 19, no. 2, pp. 73–83, 2013.

[11] T. D. Fry, R. A. Leitch, P. R. Philipoom, and Y. Tian, "Empirical analysis of cost estimation accuracy in procurement auctions," *International Journal of Business and Management*, vol. 11, no. 3, p. 1, 2016.

[12] H. Jung, G. Kosmopoulou, C. Lamarche, and R. Sicotte, "Strategic bidding and contract renegotiation," *International Economic Review*, vol. 60, no. 2, pp. 801–820, 2019.

[13] K. Bloomfield, T. Williams, C. Bovis, and Y. Merali, "Systemic risk in major public contracts," *International Journal of Forecasting*, vol. 35, no. 2, pp. 667–676, 2019.

[14] P. Ballesteros-Pérez, M. Skitmore, E. Pellicer, and J. H. Gutiérrez-Bahamondes, "Improving the estimation of probability of bidder participation in procurement auctions," *International Journal of Project Management*, vol. 34, no. 2, pp. 158–172, 2016.

[15] T. Hanák and P. Muchová, "Impact of competition on prices in public sector procurement," *Procedia Computer Science*, vol. 64, pp. 729–735, 2015.

[16] V. Titl and B. Geys, "Political donations and the allocation of public procurement contracts," *European Economic Review*, vol. 111, pp. 443–458, 2019.

[17] S. Tadelis, "Public procurement design: lessons from the private sector," *International Journal of Industrial Organization*, vol. 30, no. 3, pp. 297–302, 2012.

[18] T. Hanák and C. Serrat, "Analysis of construction auctions data in Slovak public procurement," *Advances in Civil Engineering*, vol. 2018, Article ID 9036340, 13 pages, 2018.

[19] J.-M. Kim and H. Jung, "Predicting bid prices by using machine learning methods," *Applied Economics*, vol. 51, no. 19, pp. 2011–2018, 2019.

[20] Publications Office of the European Union, *The Official Journal of the European Union*, Publications Office of the European Union, Brussels, Belgium, 2019, https://eur-lex.europa.eu/oj/direct-access.html.

[21] Publications Office of the European Union, *European Union Open Data Portal*, Publications Office of the European Union, Brussels, Belgium, 2019, http://data.europa.eu/euodp.

[22] Tenders Electronic Daily (TED), "Online version of the supplement to the official journal of the EU," 2019, https://ted.europa.eu.

[23] European Commission, Eurostat, https://ec.europa.eu/eurostat.

[24] L. Breiman, "Random forests," *Machine Learning*, vol. 45, no. 1, pp. 5–32, 2001.

[25] A. Verikas, A. Gelzinis, and M. Bacauskiene, "Mining data with random forests: a survey and results of new tests," *Pattern Recognition*, vol. 44, no. 2, pp. 330–349, 2011.

[26] M. Fernández-Delgado, M. S. Sirsat, E. Cernadas, S. Alawadi, S. Barro, and M. Febrero-Bande, "An extensive experimental survey of regression methods," *Neural Networks*, vol. 111, pp. 11–34, 2019.

[27] M. R. Segal, *Machine Learning Benchmarks and Random Forest Regression*, UCSF: Center for Bioinformatics and Molecular Biostatistics, San Francisco, CA, USA, 2004, https://escholarship.org/uc/item/35x3v9t4.

[28] G. Biau and E. Scornet, "A random forest guided tour," *Test*, vol. 25, no. 2, pp. 197–227, 2016.

[29] T. Hastie, R. Tibshirani, and J. Friedman, *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*, Springer, Berlin, Germany, 2nd edition, 2008.

[30] M. B. Kursa and W. R. Rudnicki, "Feature selection with the Boruta package," *Journal Of Statistical Software*, vol. 36, no. 11, pp. 1–13, 2010.

[31] H. Deng and G. Runger, "Gene selection with guided regularized random forest," *Pattern Recognition*, vol. 46, no. 12, pp. 3483–3489, 2013.

[32] T. Hothorn, K. Hornik, and A. Zeileis, "Unbiased recursive partitioning: a conditional inference framework," *Journal of Computational and Graphical Statistics*, vol. 15, no. 3, pp. 651–674, 2006.

[33] N. Meinshausen, "Quantile regression forests," *Journal of Machine Learning Research*, vol. 7, pp. 983–999, 2006.

[34] P. Geurts, D. Ernst, and L. Wehenkel, "Extremely randomized trees," *Machine Learning*, vol. 63, no. 1, pp. 3–42, 2006.

[35] Y. Cheng, X. Chen, X. Ding, and L. Zeng, "Optimizing location of car-sharing stations based on potential travel demand and present operation characteristics: the case of Chengdu," *Journal of Advanced Transportation*, vol. 2019, Article ID 7546303, 13 pages, 2019.

[36] Q. Shang, D. Tan, S. Gao, and L. Feng, "A hybrid method for traffic incident duration prediction using BOA-optimized random forest combined with neighborhood components analysis," *Journal of Advanced Transportation*, vol. 2019, Article ID 4202735, 11 pages, 2019.

[37] J. Xing and G. Zheng, "Stress field gradient analysis technique using lower-order C^0 elements," *Mathematical Problems in Engineering*, vol. 2015, Article ID 457046, 12 pages, 2015.

[38] Z. Sun, H. Sun, and J. Zhang, "Multistep wind speed and wind power prediction based on a predictive deep belief network and an optimized random forest," *Mathematical Problems in Engineering*, vol. 2018, no. 4, Article ID 6231745, 15 pages, 2018.

[39] Z. Liao, Y. Ju, and Q. Zou, "Prediction of G protein-coupled receptors with SVM-prot features and random forest," *Scientifica*, vol. 2016, Article ID 8309253, 10 pages, 2016.

[40] L. Dong, X. Li, and G. Xie, "Nonlinear methodologies for identifying seismic event and nuclear explosion using random forest, support vector machine, and naive bayes classification," *Abstract and Applied Analysis*, vol. 2014, Article ID 459137, 8 pages, 2014.

[41] T. Xiang, T. Li, M. Ye, and Z. Liu, "Random forest with adaptive local template for pedestrian detection," *Mathematical Problems in Engineering*, vol. 2015, Article ID 767423, 11 pages, 2015.

[42] H. R. Zhang, F. Min, and X. He, "Aggregated recommendation through random forests," *The Scientific World Journal*, vol. 2014, Article ID 649596, 11 pages, 2014.

[43] J. Ruyssinck, J. van der Herten, R. Houthooft et al., "Random survival forests for predicting the bed occupancy in the intensive care unit," *Computational and Mathematical Methods in Medicine*, vol. 2016, Article ID 7087053, 7 pages, 2016.

[44] A. A. Akinyelu and A. O. Adewumi, "Classification of phishing email using random forest machine learning technique," *Journal of Applied Mathematics*, vol. 2014, Article ID 425731, 6 pages, 2014.

[45] E. Min, J. Long, Q. Liu, J. Cui, and W. Chen, "Tr-ids: anomaly-based intrusion detection through text-convolutional neural network and random forest," *Security and Communication Networks*, vol. 2018, Article ID 4943509, 9 pages, 2018.

[46] X. Gao, J. Wen, and C. Zhang, "An improved random forest algorithm for predicting employee turnover," *Mathematical Problems in Engineering*, vol. 2019, Article ID 4140707, 12 pages, 2019.

[47] National Statistics Institute (INE), Spain, https://www.ine.es.