VILNIUS TECH
Vilnius Gediminas Technical University

Journal of Civil Engineering and Management
ISSN 1392-3730 / eISSN 1822-3605
2022 Volume 28 Issue 8: 680–695
https://doi.org/10.3846/jcem.2022.16944

BUDGET AND COST CONTINGENCY CART MODELS FOR POWER PLANT PROJECTS

Md ARIFUZZAMAN[1], Uneb GAZDER[2], Muhammad Saiful ISLAM[3*], Martin SKITMORE[4]

[1]*Department of Civil and Environmental Engineering, College of Engineering, King Faisal University, Al-Hofuf, Saudi Arabia*
[2]*Department of Civil Engineering, University of Bahrain, Isa Town, Bahrain*
[3]*School of Engineering and Technology, Central Queensland University, Melbourne, Australia*
[4]*Faculty of Society and Design, Bond University, Gold Coast, Australia*

Received 30 September 2021; accepted 30 March 2022

Abstract. Cost overruns are a ubiquitous feature of construction projects, and realistic budgeting at the development stage plays a significant role in their control. However, the application of existing models to budgeting for power plant projects is restricted by the limited amount of project-specific cost data available. This study overcomes this by using a Classification and Regression Tree (CART) approach involving mixed methods of website visits, document study, and expert opinion to predict the amount of project cost (PC) and cost contingency (CC) needed to cover probable cost increases by the use of models containing readily available project attributes and national economic parameters at the project development stage. The modeling process is demonstrated and tested with a case study involving 58 Bangladeshi power plant projects – producing average absolute errors ranging from 0.7% to 1.7% and enabling project cost, inflation rate, and GDP to be identified as significant factors affecting PC and CC modeling. The approach can be applied to predict the PC during preliminary budgeting and selecting a project type and location aligned to the country's economic status and policy-making strategies, thus facilitating further investment decisions.

Keywords: power plant, project cost, cost contingency, prediction, CART.

Introduction

Realistic budgeting at the development stage of construction projects plays a significant role in controlling cost overruns – a ubiquitous problem affecting construction work worldwide (Amadi, 2021). However, budgeting for power plant projects is restricted by a limited ability to use existing project cost (PC) prediction models because of the inadequate amount of project-specific cost data available due to their relative uniqueness on account of the nature of the investment and involvement of stakeholders (Aragonés-Beltrán et al., 2014; Zhao et al., 2019). Of the few studies to date, Jung et al. (2016) and Diab et al. (2017), for instance, consider the management of risks and cost contingency (CC) (amount needed to cover probable cost increases) of mega construction projects, but the findings of such studies make only a limited contribution to the planning and management of power plant projects due to their individualistic nature and complex construction processes. In addition, most studies focus on establishing the CC needed based on expected risks in the project execution phases (Ayub et al., 2019). Identifying and assessing the potential and critical risks in the preliminary phases is a complicated process for power plant projects due to their long duration and limited number of similar past projects, which renders the findings of existing studies less useful for those involved in their planning and development.

PC and CC models also have common limitations. For example, Monte Carlo Simulation (MCS), Artificial Neural Network (ANN), Case Based Reasoning (CBR), Support Vector Machine, and multiple/stepwise regression models require intensive and quantitative data from similar previous projects. MCS and ANN, in particular,

*Corresponding author. E-mail: *m.islam3@cqu.edu.au*

can provide more accurate cost predictions for building and road transportation projects, where historical data are available (Elmousalami, 2020a, 2020b). Of these, MCS requires the mean and standard deviation of historical cost data from similar projects to develop a probability distribution function (Chang & Ko, 2017) – a constraint that restricts the application of MCS to power plant projects, which are less frequent and with limited access to cost datasets (Islam et al., 2021). Williams and Gong (2014) use cost data of 92 building projects with an integrated ANN and support vector machine (SVM) for better PC prediction accuracy, and Dursun and Stoy (2016) use 657 building projects for PC prediction by applying an ANN-based multistep ahead approach. While ANN performs better than the multiple regression model (Hashemi et al., 2019), it involves a greater amount of trial and error. ANN provides better accuracy with the combination of a complex Genetic Algorithm or Support Vector Machine (Günaydin & Doğan, 2004). Again, however, the prediction accuracy of ANN, CBR, and SVM models is compromised if there is a limited training dataset and noisy or missing data (Elfahham, 2019). Moreover, ANN is limited in its handling of uncertainty in project execution and is a black box system where the human estimator has no control (Elmousalami, 2020b). In addition, the accuracy of these models depends on the quality of the datasets used (Barraza et al., 2007; Chang & Ko, 2017; Hammad et al., 2016; Maronati & Petrovic, 2019; Shahtaheri et al., 2016). However, it is critical to ensure the availability of both the quantity and quality of cost-data for complex power plant projects. Expert judgment-based models, such as fuzzy set theory, fuzzy expert systems, and fuzzy-Bayesian belief networks require data to be elicited from domain experts, which are sometimes vague, imprecise, subjective, and contingent on specific project characteristics (Idrus et al., 2011; Islam et al., 2021; Salah, 2015). These limitations conspire to make all these models generally impractical at the initiation and planning phase of such complex infrastructure projects as power plants.

In contrast, Classification and Regression Tree (CART) models are considered efficient for scaling large problems with smaller datasets due to their condition-based tree structures (Razi & Athappilly, 2005). This makes them suitable for the uniqueness of power plant projects and their smaller datasets. Therefore, this study develops CART models for predicting the PC and CC needed. These are demonstrated and validated by data from the Bangladesh power industry. The variables involved are either macro-level project attributes or economic parameters available from accessible sources (i.e., relevant websites, published project reports, and collected project documents/information from experts). Parameters related to the *types of power plant project* (*i.e.*, *natural gas*, *combined cycle*, *heavy fuel oil*, *etc.*), *project delivery systems*, *construction site location*, and *national economic conditions* (*inflation rate*, *total GDP*, *GDP in construction*, *etc.*) are used as inputs, as these are easier to determine and plan, which makes the analysis and models useful for both government and private entities. The unique contributions of this study are that:

- different CART models are combined to form an ensemble to predict the most critical contingent cases: ensembles such as this have not been utilized in previous work;
- the models use smaller datasets than hitherto for predicting the PC and CC needed for such complex and large infrastructures as power plants;
- the ensemble is used to predict the most critical cases with the least and highest CC needed;
- the critical factors and their values affecting PC and CC are identified.

The remainder of this paper summarizes the literature concerning the cost performance of power plant projects worldwide, the models applied or developed for their predicted PCs and CCs, research methodology, the CART model and its demonstrated application to a set of Bangladesh power plant projects, and discussion of the outcomes obtained. Finally, conclusions are drawn, policy implications identified, and recommendations made for further research and application of the models.

1. Literature review

1.1. Cost performance of power plant projects

Unexpected extra costs of building an energy infrastructure project can directly influence the price and guidelines of electricity in a viable market (Gilbert et al., 2017). Cost overruns for power plant projects have been, and continue to be, ubiquitous to the point of global crisis. For example, a study of international electrical infrastructure projects found their cost to be an average of 66% over budget regardless of their type, location, and size (Sovacool et al., 2014). Li and Wang (2018) evaluation of the cost overruns of public private partnership Chinese electricity projects determines the complexity of accessing credit from various sources, financing environmental risks, risks from frequent policy changes with a government change, and allocating risks between project parties. Eybpoosh et al. (2011) present a risk network to understand the most critical risk opportunities and their impact on the cost overruns of Turkish power plants, and identify the reasons that represent the most critical risks causing the overruns linked to entrepreneurs, such as the lack of technical, financial, and human resources.

Of other energy resources, nuclear power plant projects have the highest cost overruns in the world, followed by hydropower plants (Sovacool et al., 2014). For instance, 60% of Uganda's Hydropower Projects (HPPs) experienced around 20% cost overruns (Awojobi & Jenkins, 2016). According to Xia et al. (2017), corruption, tight schedules, government bureaucracy, government intervention, lack of competition in the supply phase, and defective design are the main risks involved. In contrast, wind power plants generally have only a slight cost overrun worldwide, although they have particular problems (such as demand-

ing a significant investment in the initial phases, site selection complexity, and environmental requirements), which make them very difficult to complete within their stipulated time (Sovacool et al., 2014). Another study that deals specifically with the external risks (e.g., ecological, economic, and socio-political) of a Turkish wind project found significant influences on costs to be changes in laws and regulations, protection of the natural area, environmental problems (adequate wind flow, conservation of areas, avoidance of main bird migration routes, etc.), restrictions on land use and building permits, and the potential impact of long transmission lines on residents.

Previous studies of the cost performance of Bangladesh power plant projects found that they struggle to hold to budgeted cost, regardless of project type, size, and contracting system – the major risks contributing to cost overrun being the owner's bureaucratic complexity, land acquisition delays, delays in project tendering and the owner's decision-making, and lack of materials and equipment available in the local market (Islam & Nepal, 2016; Islam et al., 2018, 2019).

1.2. PC and CC prediction models

The most frequently available models for predicting PC and CC are basic statistical/deterministic models (Hoseini et al., 2020), probabilistic models (Touran, 2003; Uzzafer, 2013), MCS (Barraza et al., 2007; Chang & Ko, 2017; Hammad et al., 2016; Maronati & Petrovic, 2019; Shahtaheri et al., 2016), fuzzy set theory (Jung et al., 2016; Salah & Moselhi, 2015), fuzzy expert systems (Idrus et al., 2011), fuzzy-Bayesian belief network (Islam et al., 2019), regression models (Thal et al., 2010), and artificial neural networks (ANN) (Diab et al., 2017; Lhee et al., 2012) and machine learning (Bilal & Oyedele, 2020; Chakraborty et al., 2020). A brief analysis of these models is presented below to highlight the importance of the CART model used in this study.

Hoseini et al. (2020) use basic statistical analysis (goodness-of-fit test, mean, standard deviation, distribution pattern, etc.) for establishing a CC in practice based on historical data. The CC is assigned to the known-unknown and unknown-unknown risks prior to the project execution phases. However, their approach to CC modeling depends on the cost data records in the preconstruction phases of similar previous projects. Touran (2003) presents a probabilistic cost model that considers the risks as cost variables, requires a comparatively small amount of data, and produces a moderately high prediction accuracy – a critical aspect of this model being assuming the variables to be independently and identically normally distributed. The model also provides a deterministic cost value of a risk, which is always questionable in the face of uncertainty and subjective judgment. Uzzafer (2013) also demonstrates a probabilistic CC model integrating risk assessment and management strategies in PC prediction. The model advances the combined application of experts' elicited and historical datasets for CC modeling. The challenges of applying this model are in assigning a cost value and probability to an individual risk, assuming no inter-relationships between risks, and considering the worst case of the risk only.

MCS, an advanced probabilistic model, is a powerful tool for PC prediction and CC modeling in uncertain and complex project environments, and has been applied in predicting the PC or CC in many studies. Barraza et al. (2007), for instance, apply it to predicting the contingency of each activity of a project considering its probability distribution under different cases, and taking into account the risk management strategies for modeling each activity cost. Shahtaheri et al. (2017) integrate MCS with the risk assessment approach for CC prediction. However, the model requires sufficient reliable historical data to generate a probability density function for better prediction accuracy. Moreover, its basic assumption is that the cost variables are discrete, independent, and normally distributed – the independent assumption being unrealistic for such a complex project as a power plant. Hammad et al. (2016) use an MCS model to predict the CC for each activity, denoting importance to its percentage contribution to total PC in predicting the CC of a whole project as the difference between total and the planned cost. This model significantly overcomes the limitation of a data-intensive MCS model as it can accommodate expert judgment-based predictions in the absence of historical data from previous similar activities/projects. In a similar study, Maronati and Petrovic (2019) use MCS to simulate individual cost variables (cost of work, price of materials, and equipment) separately and then combine them to model total PC by the distribution-free rank correlation between the cost variables. Mawlana and Hammad (2015) quantify the impact of uncertainty, correlation between the events, and simulate the project cost and contingency cost using join probability theory. Their model considers all possible scenarios of an event and simulates the combination of multiple events finding the best possible combination of PC and schedule. However, the limitation of the number of factors to consider in a joint probability is a point of argument to make the model a practical tool for project cost or contingency budgeting.

Such other expert judgment-based methods as fuzzy set theory, fuzzy expert systems, and fuzzy-Bayesian networks are commonly used for modeling PC and CC integrated with risk and uncertainty assessment and management in a complex project environment. For instance, Salah and Moselhi (2015) developed a contingency depletion curve for monitoring and controlling contingency funds over the project execution phases. In this model, an expert can express the CC using fuzzy numerical functions considering the level of risk/uncertainty, and the predicted CC is allocated to individual work packages for its better management. The fuzzy set theory-based model does not require any historical cost data and reduces the significant amount of time needed for CC computation by simulation-based models. However, the use of expert judgments instead of economic parameters (inflation, GDP, etc.) is

subjective, biased, and vague, which can significantly reduce cost prediction accuracy. Jung et al. (2016) address this limitation of fuzzy set theory-based CC modeling by integrating and assessing the risks involved in PC overruns; however, their model does not account for any interrelationships between risks, and disregards their dynamic behavior in different project phases. Idrus et al. (2011) present a fuzzy expert system, which uses the frequency and severity of risks directly as inputs for predicting CCs, and can accommodate the contractor's experience and judgment in CC prediction: its limitations are in disregarding the interrelationships between risks, summing the magnitudes of all risks for computing CC, and ignoring risk management strategies in CC allocation. The activities or cost items of complex project infrastructure projects are also highly correlated, which cannot be handled by fuzzy set theory. Accordingly, Islam et al. (2021) present a fuzzy-Bayesian belief networks (fuzzy-BBNs) approach for risk-induced CC modeling, which can handle interrelationships between the risks, and risk dynamism. However, the developed fuzzy-based models depend on expert judgment-based datasets, which are subjective, imprecise, and vague, while the collection of subjective datasets at the early stage for project budgeting is also a complex task.

The multiple linear regression model has commonly been applied for the PC and CC modeling of different construction projects. Thal et al. (2010), for example, analyze cost data available from similar previous projects for CC prediction immediately prior to the contract award, which is an advantage over the traditional fixed percentage-based CC allocation approach. However, the model does not address the potential risks and uncertainties involved, and some qualitative variables that have substantial cost effects are not considered because of lack of data. Diab et al. (2017) present an integrated approach that includes a relative importance index and stepwise regression model. The model calculates the expert judgment-based relative importance of identified risks for CC estimates. It establishes the interrelationship between risks and subsequent costs to estimate the CC needed. The approach is deterministic and useful for allocating CC at the planning stage. The advantage of the model is that it identifies the major risks involved and considers CC prediction as simply a function of the relative importance of the risks. The model finally develops a first-order equation for estimating CCs, in which risks are considered as variables. The ANN model – an advanced form of regression analysis – is adopted by Lhee et al. (2012). This can work with many input variables, including project characteristics, to predict a single PC or CC from the history of previous projects (contract value and actual value), predict CCs as accurately as possible, and assist project administrators in realistic CC prediction and project funding. The accuracy of the ANN model is controlled by an optimal number of hidden layers, which is fully controlled by the user, and depends on access to reliable historical data from similar previous projects: more data can provide better accuracy. Chakraborty et al. (2020) compare the performances of some machine learning (ML) models such as ANN, random forest, light gradient boosting, natural gradient boosting, extreme gradient boosting, and linear regression based on their estimation accuracy, uncertainty, and time requirement to perform the job. They also use a game theory interpreting a model's performance. Accordingly, natural and light gradient boosting models are found efficient to model project costs. The mechanisms of machine learning models are usually unknown to many professionals as ML models are mostly having uncontrolled or unexplained black-box. Thus, Bilal and Oyedele (2020) propose a specific guideline for the applications of ML models in project cost estimation and demonstrate their model to estimate a project's profit margin, which is critical to winning a bid.

The CART model, on the other hand, provides an alternative means of overcoming the problem of the ANN model's uncontrolled hidden layers as it generates logical and explainable relationships between the variables. It is also useful for finding important categorical parameters (Elmousalami, 2020a; Perner et al., 2001), provides high-performance computational efficacy by splitting the parameters, and overcomes the limitation of dimensionality of the variables (Prasad et al., 2006). However, it performs poorly with nonlinear and time-series data (Curram & Mingers, 2017). The potential application of the CART model for predicting construction or CC in a complex infrastructure project environment is as yet unresearched (Elmousalami, 2020a).

2. Research methodology

This study uses CART for predicting a power plant PC and CC using real-life project characteristics and national economic factors. Several cost prediction models are developed and the factors' strengths for CC prediction are established through ensembles. The following subsections briefly describe the data collection process and characteristics of the dataset, data organization and presentation for developing cost prediction models, and introduce the CART model with a comparative discussion of its advantages over other similar methods.

2.1. Data collection

Ten variables (Table 1) are considered for modeling the cost and contingencies of power plant projects in Bangladesh. The project-related variables are construction cost, contingency in construction cost, location of the power plant, owner organization of power plant, plant type based on the power generation mechanism, type of contract under which the project is constructed, and the power generation capacity of the plant; and the economic indicators include total gross domestic product (GDP) of country, construction GDP, and inflation rate. Previous similar studies advise considering these factors for cost or CC modeling at the project's development stages. For example, Hashemi et al. (2019) consider project power plant type,

Table 1. List of variables used for modeling

Variable	Description
Cost	Construction cost of project (USD million)
Cont.	Estimated CC (% of cost)
GDP	Total Bangladesh GDP for the year of construction of the project (USD million)
Const. GDP	Bangladesh construction sector GDP for the year of construction of project (USD million)
IR	Inflation rate for Bangladesh for the year of construction of the project (%)
Location	Location of construction, divided into three categories: Urban, Rural, and Peri-Urban
Owner	An organization that is the owner of the project, divided into three categories: Government, Independent Power Producer (IPP), and Semi-autonomous
Plant	Type of power plant, divided into four categories: Combined Cycle Power Plant (CCPP), Coal, Heavy Fuel Oil (HFO), and Natural Gas
Contract	Type of construction contract award system divided into three categories: Engineering, Procurement, Construction (EPC), Build-Own-Operate (BOO), and Turnkey
Power	Power generation capacity of the project (MW)

project duration, project phase, etc., as input variables for conceptual cost estimation for power plant projects, while other studies (Elfahham, 2019; Islam et al., 2017; Lam & Siwingwa, 2017; Musarat et al., 2021) examine the inflation rate in project budgeting. Lee et al. (2017) study such construction risks as project type, contract type, project location, inflation rate, interest rate, and gross national income to predict the management reserve for international projects.

The project-related data are collected from three sources: (1) visiting the websites of the Bangladesh Power Development Board (BPDB), Northwest Power Company Limited (NWPCL), and Ashuganj Power Company Limited (APCL); (2) study of project documents; and (3) expert judgment. The data relating to construction year, plant ownership (BPDB, NWPCL, APCL, etc.), location (urban, peri-urban, and rural), plant type (Combined Cycle Power Plant (CCPP), Heavy Fuel Oil (HFO), Coal, and Natural Gas), power generation capacity (Megawatt), type of contract (Engineering, Procurement, and Construction (EPC), Build-Own-Operate (BOO), and Turn-Key), and estimated budget for the projects are found from the corresponding organizations' websites. However, estimated CC (%) and cost overrun (%) are obtained by studying the collected project documents and expert judgment.

The experts were contacted following different approaches (i.e., by email, telephone call, or direct visit by appointment) to collect project documents and missing data. The email list and contact details of plant managers or project directors were collected from their affiliated organization's website, member list of the Institute of Engineers Bangladesh (IEB), and directly visiting country-wide project sites and their corporate offices. With these approaches, approximately 60 randomly selected projects were targeted, and 120 experts (two per project) were requested to provide project data based on their documents or judgments. Of these, 73 experts responded with complete information of 58 projects – a response rate of 60.83%, which is deemed acceptable compared with previous studies (Ajay & Micah, 2014; Maas & Hox, 2005; Olaniran, 2015; Singh & Masuku, 2013). A report published in 2020 shows that Bangladesh has 138 power plant projects with the highest 12,983 Megawatt (MW) power generation capacity (Haque, 2020). Thus, the collected projects represent 42% of the total constructed projects in Bangladesh, which justifies using statistical inference (Olaniran, 2015). In similar studies, Hashemi et al. (2019) use data for 39 projects, and Gunduz and Sahin (2015) use data for 54 projects. The practical challenges to the data collection were the unwillingness of the experts, fear of disclosing project information, and lack of quality data records. The data relating to the country's economic indicators, such as total GDP, construction GDP, and inflation rate, were collected from the International Monetary Fund [IMF]. The IMF is an international organization of 190 countries monitoring and funding sustainable economic growth, working for poverty alleviation and financial stability worldwide (IMF, 2020). The data were cross-checked with the Bangladesh Bureau of Statistics [BBS] (2020).

2.2. Data organization and presentation

The collected data are organized against ten categorical variables as presented in the Appendix (Table A.1). Of the 58 projects involved, the earliest project was constructed in 1993, while the latest started in 2020. The average PC is approximately USD 290 million, while the average power generation capacity is approximately 258 MW. Each of the ten categorical variables was further divided into binary variables. Table 2 provides the percentage distribution of the projects in terms of binary variables related to ownership, plant type, type of contract, and project location. For instance, 53% of the projects are in rural areas, 40% in peri-urban areas, and the remaining in urban areas. In addition, most projects, 43%, are owned by the BPDB, 57% are CCPP, while 76% were awarded through EPC contracts.

Table 2. Distribution of projects in terms of ownership, plant type, and contract type

Project variables	Binary variables	Proportion
Project ownership organizations	BPDP	43%
	Semi-autonomous organization	33%
	IPP	24%
Plant type	CCPP	57%
	HFO	21%
	Natural gas	16%
	Coal	6%
Contract type	EPC	76%
	BOO	12%
	Turnkey	12%
Locations	Rural	53%
	Peri-urban	40%
	Urban	7%

The inflation rate, total GDP, and construction GDP of Bangladesh are available until 2019. The forecasted values for inflation rate and total GDP are also reported up to 2024 by the IMF (2020). Figure 1 shows the trend of these values, there being similar trends for total GDP and construction GDP with rapid growth in recent years, while the inflation rate fluctuates wildly unto 2012.

The average estimated CC is 7.2%, with a standard deviation of 2.6%, and the minimum and maximum CC estimates are 1% and 10%, respectively. For large construction projects, the estimated proportion of CC depends on the construction cost of the project (Thal et al., 2010).

2.3. CART models

CART is an unsupervised artificial intelligence technique. It has adaptive interpretation skills and can successfully handle complex non-linearities between predictor and response variables and multi-collinearity problems of the data better than regression models (Gong et al., 2018). Unlike other artificial intelligence and machine learning tools, CART makes no such assumptions as the data should be random, independent, have linear or nonlinear relationships, etc. This is one of the significant advantages of using CART to develop a cost prediction model. In addition, the CART analysis provides a model that can be interpreted through logical statements to understand the effect of different variables on the target variable, which is rarely possible with other data mining tools (Shaaban & Pande, 2016).

Two CART models are developed in this study: firstly, for predicting the CC, and then for predicting the PC. These models have gained popularity during the past five decades (Loh, 2014) since their inception as a non-parametric modeling technique, especially for handling ecological data (Moisen, 2008; Steinberg, 2009). These models utilize historical data to construct decision trees by calculating the optimal distribution/separation of output data with respect to ranges of input variables (Timofeev, 2004). An appropriate set (division) of variables is found at each node of the decision tree, which minimizes the error.

The process is continued until no significant improvement can be achieved with the further division of a node, which then becomes a terminal node for the tree (Strobl et al., 2009). These models work by partitioning the predictor space into rectangles, which is based on rules used to identify regions with the most homogeneous responses to predictors. Then, a constant is fitted to each region (Figures 2 and 5), with trees fitting the mean response for observations in that region. The errors are assumed to be normally distributed (Elith et al., 2008). The model development follows the following process:

(1) Each independent variable in the dataset is taken as a node and a split is found that causes maximum variance in the dependent variable – a procedure often termed a "recursive binary split".
(2) The split is used to generate child nodes for that node, the above step being repeated for the independent variables in the dataset.
(3) When a node does not show any split of values where variance is possible, it becomes a terminal node.
(4) The above steps are repeated for all the child nodes with the precondition of the parent nodes.

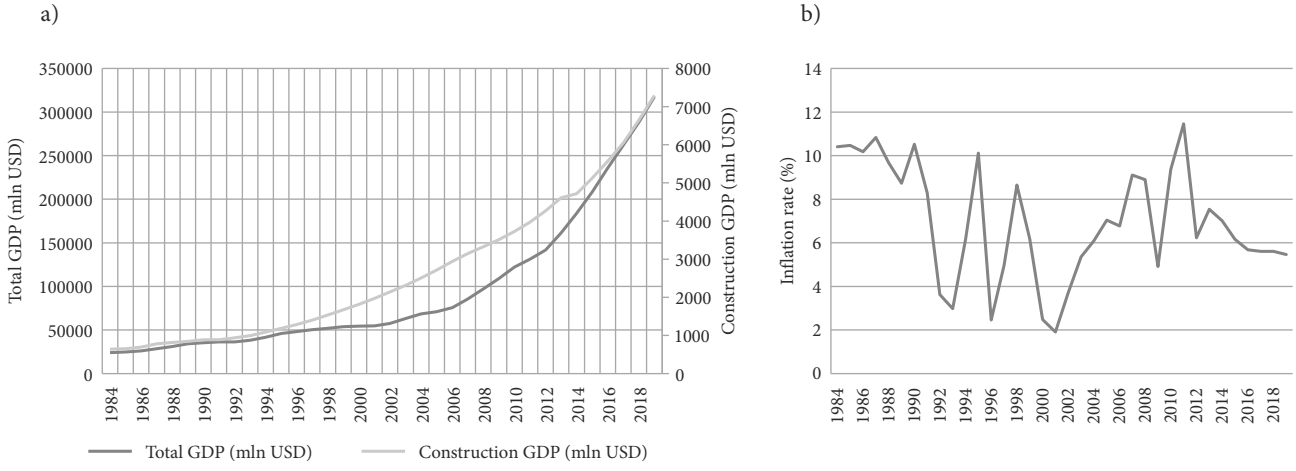

Figure 1. a – Bangladesh's GDP trend, and b – Bangladesh inflation rate trend

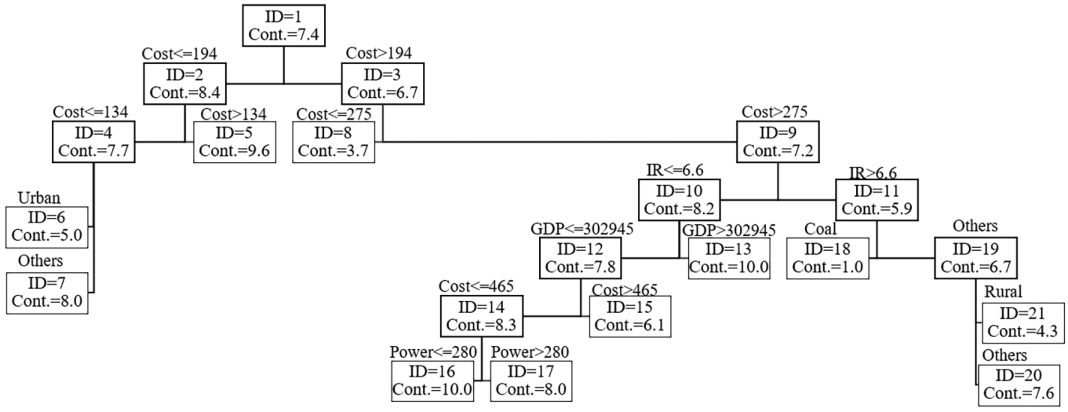

Figure 2. CART for CC modeling of the power plant projects

3. Model development and results

3.1. Cost prediction using the CART model

All the information for each of the 58 projects is tabulated in the Appendix (Table A.1). The datasets are divided into 37 training projects and 21 testing projects, following Hashemi et al. (2019), where they use 60% of the data for training and the remainder for cross-validation and testing the prediction accuracy of the trained model. The projects in each dataset are selected based on their variety. The training and testing datasets have projects from all possible combinations of project location, ownership organization, and plant and contract types. The data are classified according to the project location, owner organization, plant type, and contract type. Then, projects having the same response for all the above variables are organized. Afterward, approximately 66% of the projects are assigned to the training dataset, and the remaining for the test dataset. For example, the training dataset has 66% of the CCPP projects built in an urban area, owned by a government organization, and awarded by the EPC contract. The remaining 34% is assigned to the test dataset.

Figure 2 shows the developed CART model for CC prediction. PC is the most important variable, which affects the tree at several levels, including the top node. Project type, location, GDP, power generation, and inflation rate are the other variables that affect the model. Node 18 has the least CC (1%), which is a coal power plant costing over USD 275 million with an inflation rate of over 6.6%. Nodes 13 and 16 have the highest CC. For Node 13, the PC is over USD 275 million with an inflation rate of 6.6% or less and a GDP of over USD 302,945 million. For Node 16, the PC is between USD 275 and 465 million, with an inflation rate of again 6.6% or less, a GDP of USD 302,945 million or less, and the power generation capacity of 280 MW. Projects with PC between USD 134 and 194 million also have high CC estimates (see Node 5) irrespective of any other factor, while projects in urban areas have lower CC estimates than in other locations (see Nodes 6 and 7).

Figures 3 and 4 are the scatterplots for the estimated CC value and model error (observed minus predicted) for the training and validation datasets, respectively. The

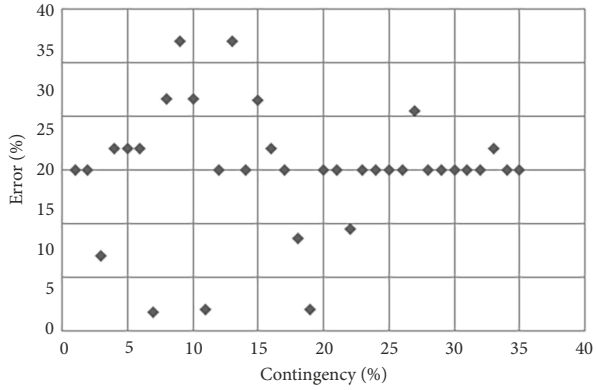

Figure 3. Scatterplot of the CART CC training dataset predictions

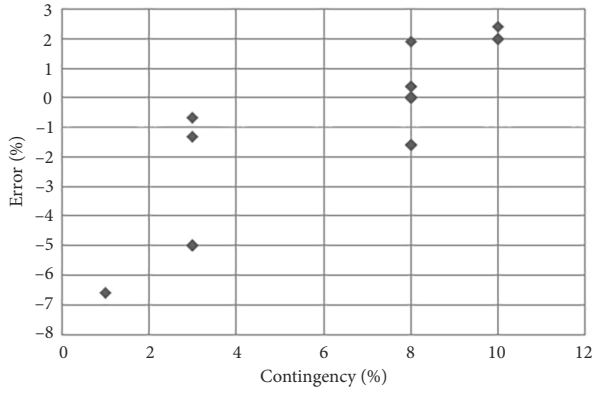

Figure 4. Scatterplot for the CART CC testing dataset predictions

Average Absolute Error (AAE) for the model is 0.7% and 1.7% for the training and testing datasets, respectively.

Construction cost is the most important parameter for this CART model for CC prediction (later referred to as CART 1), which is consistent with the literature for large construction projects (Thal et al., 2010). The construction cost depends on factors related to the project attributes and the national economy, including those used from Table 2. Therefore, another CART model (Figure 5) is developed to predict the project's construction cost (later referred to as CART 2). This model shows that power generation ca-

pacity has a multilevel effect on construction cost, which seems to increase with capacity. Other important parameters are GDP and the inflation rate, which affect the cost of projects having a power generation capacity of over 188 MW. The GDP indepfendently has a multilevel effect on the cost, with higher GDP associated with a higher construction cost. This finding ensures that when the national economic conditions are better, larger investments are made in the power generation sector, and projects are carried out with higher capacity (and consequently higher cost). At a higher GDP (more than USD 196,165 million), indicating a period of high economic growth, the inflation rate seems to be inversely proportional to cost, while it is directly proportional when GDP is between USD 196,165 and 126,560 million. The type of project owner organization and location affects the cost of projects with a power generation capacity between 70 and 188 MW.

Figure 5 shows that the highest cost is at Node 17, with a power generation of over 188 MW, GDP ranging between USD 126,560 and 195,165 million, and the inflation rate over 9.5%. The minimum cost is at Node 9, with power generation between 70 and 125 MW, and the projects in urban or rural areas. Projects owned by Independent Power Producers (IPPs) have a lower cost than other organizations regardless of any other factors (shown in Nodes 10 and 11).

The PCs are predicted for training and testing datasets based on CART 2 (Figure 5). The scatterplots for both training and testing datasets are shown in Figures 6 and 7, respectively: their average absolute errors (AAE) are USD 19 and 184 million, respectively.

3.2. Ensembles

The CART models show that the prediction of CC depends upon the construction cost – one of the independent parameters in the model for estimating CC, which is, in turn, dependent on other common factors. The relationship be-

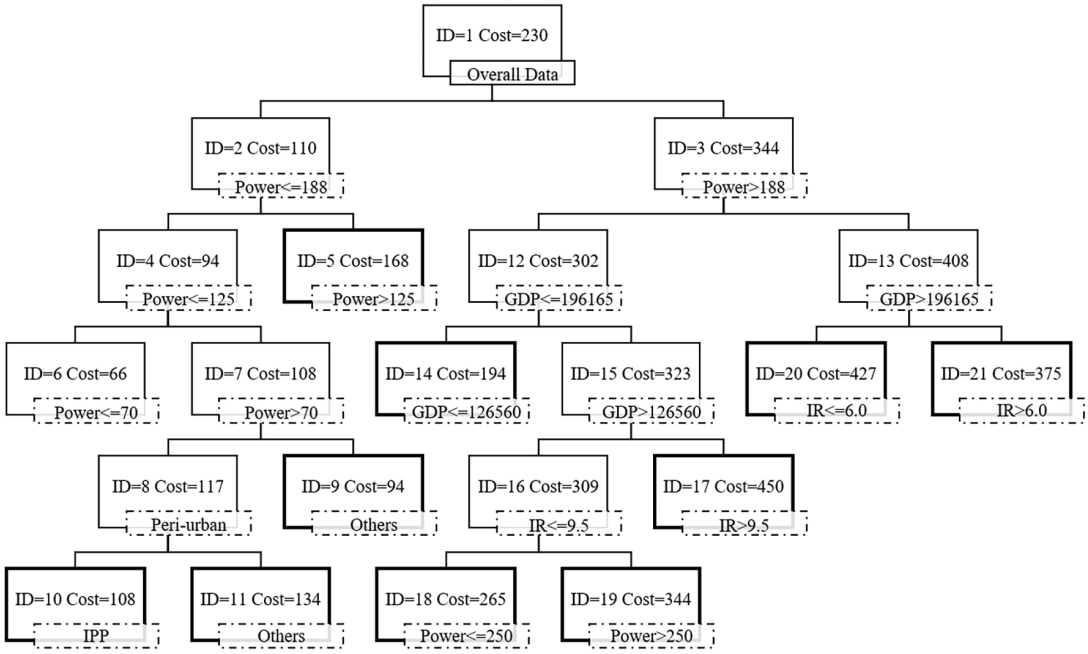

Figure 5. CART model for predicting construction costs

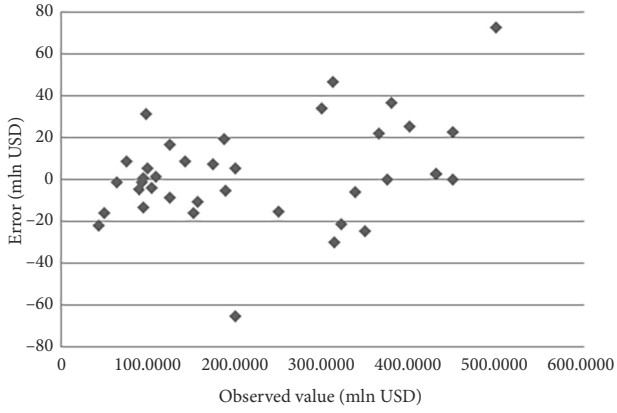

Figure 6. Scatterplot for CART PC development dataset predictions

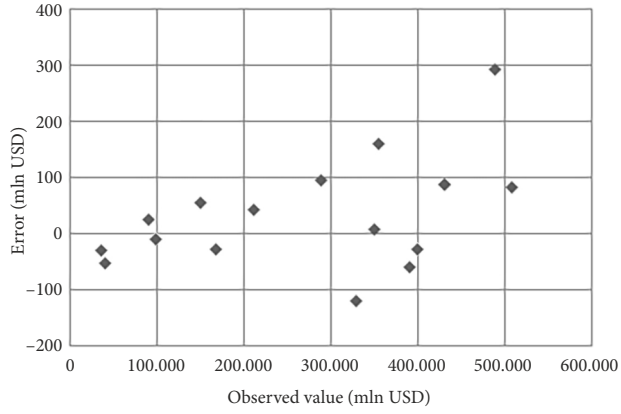

Figure 7. Scatterplot for CART PC test dataset predictions

tween the CC and PC of projects is expected and obvious. The advantage of the CART model is the elaboration of this relationship, which is nonlinear, with a decision tree structure that can determine the rules for increasing or decreasing the CC. The rules developed from both CART models (Figures 2 and 5) are combined to obtain to the nodes with the highest and lowest CC in Figure 2. These combinations are referred to as *ensembles* in this study – created for the least and highest CC paths from both models, as shown in Figures 8 and 9. Although similar ensembles can be created for all the terminal nodes of CART 1, they are omitted here to avoid further complexity for readers and decision-makers.

The least CC is found at Node 18 in CART 1 (Figure 2) for coal power plants with PCs over USD 275 million. This cost can be attained at Nodes 19, 17, and 21 in CART 2 (Figure 8). Node 20 is not included in this CART as it has an IR of 6.6% or less, which is against the condition of Node 18 in CART 1. For all possible nodes from CART 2 with a cost of more than USD 275 million, the power generation capacity is over 188 MW. If GDP is more than USD 196,165 million, then any inflation rate above 6.6 gives the least CC. However, the highest CC is at Node 17 in CART 2.

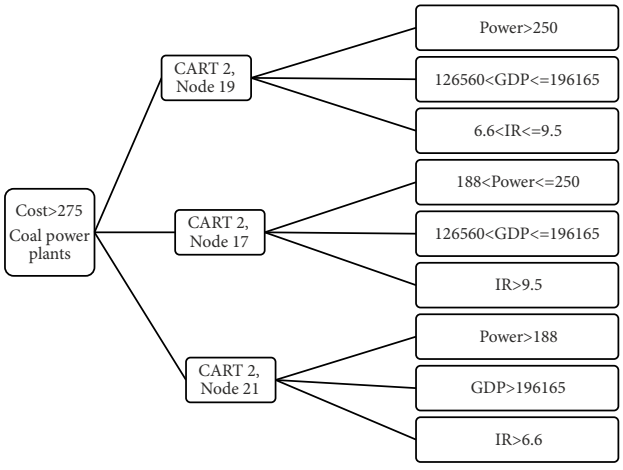

Figure 8. Least CC path ensemble

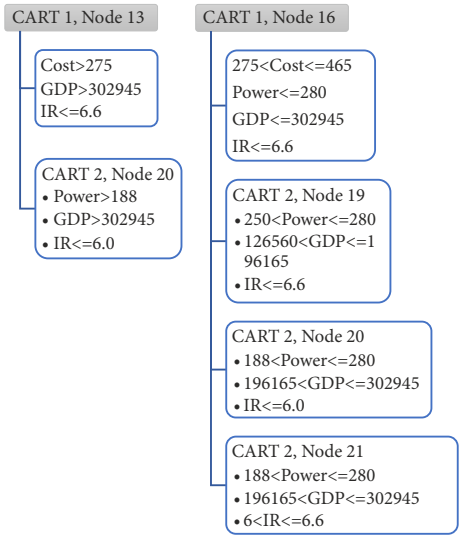

Figure 9. Highest CC path ensemble

As shown in Figure 9, the highest CC path can be achieved at Nodes 13 and 16 in CART 1. Node 13 has a cost of over USD 275 million and a GDP of over USD 302,945 million. These values can be achieved at Node 20 of CART 2. In which case, the generation capacity of the power plant should be over 188 MW, and inflation should be limited to 6.0% or less. In the case of Node 13 in CART 1, Nodes 19, 20, and 21 from CART 2 can provide the required output cost. Node 17 is not used in this case since its inflation rate is over 9.5%. If the GDP is between USD 126,560 and 196,165 million, with an inflation rate of 6.6% or less, then a 250 to 280 MW project will give the highest CC. In addition, if the GDP is between USD 196,165 and 302,945 million with an inflation rate of less than 6.6%, then any project with power generation capacity between 188 to 280 MW will give the highest CC.

The trends between the least and highest CC paths, which are shown in Figures 8 and 9 indicate that projects costing more than USD 275 million and planned for generating more than 188 MW can result in the highest or least CC. This mainly depends on national economic conditions, provided the projects are not coal-based power plants. The latter has the least CC, which may be because coal power plants have been used more commonly; hence, they have fewer associated risks. On the other hand, a lower inflation rate (less than 6.6%) will probably have higher CC estimates. This could be attributed to the lower risk of significant changes in the prices of materials and equipment resulting from lower inflation rates.

4. Discussion of the findings

The article develops new CART models and demonstrates their applications to such complex infrastructure projects as power plants, where cost variables are highly interrelated. It starts addressing some limitations of the generally available models for predicting the PC and CC of power plant projects in the introduction and the literature review sections. Existing models have limited applications as they demand a comprehensive risk assessment, which involves making a significant effort to elicit and compute experts' judgments for estimating and budgeting at the early stages of a project. For instance, Islam's et al. (2021) use of fuzzy-BBN models for Bangladesh power plant projects requires expert judgment to address uncertainty and risk assessment, which is not only impractical and time-consuming at the project development stage, but also is subjective, vague, and imprecise, which makes the value of the models' outcomes or performances questionable. The application of ANN-based models (Gunduz & Sahin, 2015; Hashemi et al., 2019), on the other hand, is restricted by the limited project-specific cost data available and uncontrolled hidden layers in the ANN "black box". Furthermore, previous studies do not always consider the project's macro-level attributes (project type, size, location, power

generation capacity, etc.) and such national economic parameters as GDP and the inflation rate. These macro-level project attributes and national economic parameters are significant considerations for risk analysis-based realistic cost estimation and contingency modelling in infrastructure projects at their early stages, i.e., the planning and design phases (Bhargava et al., 2017; Bordat et al., 2004; Gazder et al., 2020; Musarat et al., 2021). In contrast, the CART models developed in the present study utilize objective data, which include project attributes and national economic parameters as suggested in previous studies (Bhargava et al., 2017; Gazder et al., 2020; Musarat et al., 2021). The strength of the model is in its application of objective data sets that are comparatively easy to obtain, as these are readily available on the internet and other accessible sources, and thus can be easily applied for future project budgeting.

The study demonstrates the application of the model to predict the cost and contingency of different types and sizes of power plant projects under the varying economic situations (i.e., GDP and inflation rate) of a country like Bangladesh. In terms of accuracy, the developed CART model for CC prediction provides an outcome at least comparable to previously developed models. The CC prediction error (i.e., AAE) for the training dataset is 0.7%, and 1.7% for the test dataset, which are considered reasonable (Idrus et al., 2011; Islam et al., 2021; Sonmez et al., 2007) compared to similar previous studies (Gunduz & Sahin, 2015; Hashemi et al., 2019). For example, Gunduz and Sahin (2015) apply ANN and multiple regression cost estimation models to early-stage investment decision-making for 54 Turkish hydroelectric power plant projects involving 13 very project-specific variables (but without national economic parameters) – this produced a mean absolute percentage error (MAPE) of –7% to +5% for ANN and 10% for multiple regression model. Similarly, Hashemi et al.'s (2019) integrated Genetic Algorithm and ANN-based cost prediction model for 39 Iranian power plant projects involves variables of project type (i.e., gas turbine, CCPP, hydroelectric, combined gas turbine and steam, etc.) and duration, construction phases, site geology, substation construction, cooling system type, and fuel type (oil, gas, or both). This produced a predicted cost mean square error (MSE) of approximately 6%.

However, Islam's et al. (2021) work is the most relevant to the present study, with major risks and project type (CCPP, HFO, Coal, Natural Gas) and size (MW) as variables, and has a prediction error ranging from –4% to 20% – the lowest predicted CC being 8.28% for CCPP 350-400 MW projects, and the lowest prediction error of –4.84% for HFO 100 MW projects. While this is different to the present study's finding that coal-based plants should be assigned the lowest percentage of CC, it can be basically attributed to our additional inclusion of national economic parameters (i.e., national GDP, construction GDP, and inflation rate).

Another similar study of Bangladeshi power plant projects by Gazder et al. (2020) conducts a parametric cost analysis and develops a cost prediction model to show that PC varies significantly with plant type and size, and varies moderately with national GDP and construction GDP, with CCPP projects incurring the highest PC compared to other projects. Other studied also found that PC considerably varies with a project's macro-level attributes and country's general economy (Bhargava et al., 2017; Bordat et al., 2004; Hashemi et al., 2019). Moreover, the present study finds that power generation capacity or project size (MW) has a multilevel effect on the PC prediction, and construction GDP is highly correlated with PC: for instance, a higher construction GDP means a higher PC and *vice versa*.

The present study also identifies project ownership (public-funded, private, or public private partnerships) and site location as having a significant impact on PC for low power generation capacity projects (i.e., 70 to 188 MW) – an aspect not analyzed in any previous studies.

5. Implications of the CART models in future projects

Conceptual PC estimation and CC prediction models with project and national-specific parameters are important, as less than 2% of project information is available at the early stage of a project (Elmousalami, 2020b; Hegazy & Ayed, 1998). Accordingly, readily available macro-level project attributes and the country's macroeconomic parameters are capitalized as cost variables to develop CART models for predicting PC and CC. The developed CART models can be implemented for actual power plant project budgeting and financing as the models are developed solely with the data collected from the Bangladesh power plant industry and the country's economic parameters. In the project development stage in Bangladesh, a Development Project Profile (DPP) is prepared following the country's power sector development plan. Many micro-level cost variables such as the quantity and quality of plant machinery, geological properties, land acquisition and development, and amount of raw materials required for the plant's infrastructure development are unknown at the time of DPP preparation. Instead, the DPP mostly depends on type, size, proposed location, contract type, etc., along with the government's infrastructure development policy and economic parameters (Islam et al., 2018). Thus, for preliminary budgeting or DPP preparation, when risk and cost data are not available and eliciting subjective judgement is time consuming and costly, the developed CART models will be efficient for the estimators or DPP preparation team to estimate a realistic budget for a future power plant project. Thus, the developed CART models are potentially efficient and handy tools for the Bangladesh Power Development Board and other organizations involved in preparing DPPs. Moreover, the study identifies some critical relationships between the project types, sizes, location, inflation rate etc., with project cost and contingency budgeting. These findings will directly guide the BPDB authority and associated Bangladeshi ministries (i.e., The

Ministry of Power and Energy and Ministry of Finance and Planning) involved at the policy and planning level to justify and approve future projects with their budgets and locations (Bhargava et al., 2017; Bordat et al., 2004).

However, it is necessary for cost estimators to understand how to build and interpret the CART models and ensembles. Firstly, they need to locate the project-specific information and country's economic parameters needed. For example, suppose a project is to be constructed in a rural area, with a coal fuel source, planned power generation capacity of 400 MW, an EPC contract, a 2024 construction year, projected GDP of USD 470 billion, and a 5.9% inflation rate according to the BBS (2020). This condition satisfies Node 13 in Figure 2. Thus, for this specific project, the overall predicted budget is USD 427 million and CC is a maximum 10% over the budget (Figure 5, Node 20). The developed ensembles further assist in predicting a CC percentage range (i.e., highest and lowest CCs) for a project (Figures 8 and 9) as it depends on the PC having a nonlinear relationship with inflation rate, GDP, power generation capacity, and power sources. For example, coal-based power plant projects have a history of having the lowest CC (Figure 8), but the inflation rate should be greater than 6%, which does not satisfy the criteria of the above example project as the 2024 inflation rate will be less than 6%. Thus, these findings assist decision-makers in choosing a CC percentage based on their educated guess and predicted cost to find the total budget. Afterward, the policymaker can approve the budget, order a budget revision, or reject the project for financing according to the country's energy development plan and economic status. Thus, these ensembles (Figures 8 and 9) practically assist informed decision-making for PC and CC allocation of power plant projects in Bangladesh, which is a critical task for such complex and uncertain projects as power plants in their early stages.

Secondly, developing CART models for PC and CC predictions for future projects in Bangladesh or other economically similar countries can be adopted with the same or additional variables for its broader application. For this purpose, the estimator/project development team will gather such previous project data as project location, type, size, contract type, actual cost, and estimated CC, and national economic factors such as total GDP, construction GDP, and the inflation rate. Then, they will input these data to develop the model as presented in this study to obtain the predicted PC and CC for their specific type and size of the project.

Conclusions and recommendations

Power plant projects are invariably complex and uncertain. Hence, having sufficiently accurate PC and CC predictions is critical in their early stages when little project information is available. While few studies attempt to develop models for predicting power plant PC and CC, their models mainly depend on very project-specific cost datasets and expert judgment-based subjective datasets. However, access to detailed project-specific cost data is challenging due to the lack of quality data records and information-sharing policies of the project's organization. Moreover, subjective cost data is imprecise, vague, and biased on the competence of experts (experience, education, understanding of the subject matter, etc.). On the other hand, such project-specific historical information as project type, size, contract type, location, owner organization, estimated cost, and final cost, together with such national economic parameters as GDP and inflation rate, are readily available in open sources (webpages, published reports, print media, etc.). This study considers these variables for developing PC and CC models, which are partially ignored in previous studies. Accordingly, the CART models are developed here for predicting the PC and CC of power plant projects. Different CART models are then combined to form an ensemble to predict the most critical cases for contingencies in obtaining the lowest and highest amounts for CC allocation. CART-based ensembles have not been developed by any previous studies of the CC prediction of power plant projects. Another advantage of the CART and ensemble is that they can handle a smaller dataset for predicting the PC and CC of such complex and large infrastructures as power plants. Access to PC and CC data associated with project attributes and national economic conditions is much easier for the planning and budgeting departments of BPDB, NWPCL, and APCL. Moreover, the CART models' errors in predicting PC and CC for the Bangladesh power plant projects (i.e., AAE of 0.7% and 1.7% for the training and testing datasets, respectively) are less than those for other methods reported in the literature.

The study findings provide additional knowledge to the professionals and policymakers associated with the Bangladesh power plant industry. For example, the CART model shows that construction cost is the most important parameter in estimating CC, coal power plants with higher costs (more than USD 275 million) have the least CC estimate, and higher costs are associated with projects with more than 188 MW power generation capacity. The inflation rate and GDP have significant multilevel effects on both CART models despite variations in GDP – an inflation rate of less than 6.6% is associated with higher CC estimates, which could be attributed to more confidence in material and equipment prices. Projects in urban areas have a lower CC value, although the decision to construct power generation projects in urban areas is a policymaking issue as this may conflict with the living environment of residential zones. Finally, IPP projects have the lowest cost of all types of ownership. All these findings are significant for developing DPP for future projects and help guide policymakers in choosing the most suitable locations, project types, and sizes along with national economic conditions associated with huge capital investments in this critical sector.

The scope of this study was limited to the Bangladesh power industry, and limited data (58 out of 138 projects) was collected to validate the developed models. However,

CART assumes some rules to classify/split the variables based on their characteristics and inherent relationships with the dependent variables**.** This means that the developed model will produce inflated errors if the formation of rules to split independent variables is not close to the inherent relationships between independent and dependent variables because of the characteristics of the data sets. Thus, government entities such as BPDB's team need to replicate the models with larger datasets for a comprehensive justification and robustness as they own the projects' cost history and other project attributes before the practical application of the models. Other models, such as ANN or MCS, could be applied for PC and CC predictions and validated for this dataset, with their outcomes compared with the CART models' results in finding more accurate models. Although the CART models are very efficient in highlighting the optimal splitting of variables and their importance, they are not useful for sensitivity analysis because of their strict rule structure. Alternatively, the models mentioned above (ANN, MCS, etc.) can be utilized for sensitivity analysis, finding the critical variables involved in predicting a project's cost and contingency. The models used in this study, can be further utilized to develop ensembles for all terminal nodes, in addition nodes with highest and lowest CC. Additionally, further study can be conducted to develop and validate the CART models for PC and CC predictions for other infrastructure projects with limited cost data, such as bridges, airports, and power transmission and distribution projects. Furthermore, the CART modes in this study currently make some simplifying assumptions relating to the effect of construction methods, organizational culture, and project management structure on project cost and contingency predictions. Thus, future studies can consider these issues in further detail.

Data availability statement

Some or all data, models, or code that support the findings of this study are available from the corresponding author upon reasonable request.

Acknowledgements

This study was supported by the Deanship of Scientific Research, King Faisal University, Saudi Arabia [grant number GRANT1433]. The authors also acknowledge the research facilities and technical support of their affiliated universities to support this study.

References

Ajay, S., & Micah, B. (2014). Sampling techniques & determination of sample size in applied statistics research: An overview. *International Journal of Economics, Commerce and Management*, *2*(11), 1–22.

Amadi, A. I. (2021). Towards methodological adventure in cost overrun research: linking process and product. *International Journal of Construction Management*. https://doi.org/10.1080/15623599.2021.1894632

Aragonés-Beltrán, P., Chaparro-Gonzalez, F., Pastor-Ferrando, J.-P., & Pla-Rubio, A. (2014). An AHP/ANP-based multi-criteria decision approach for the selection of solar thermal power plant investment projects. *Energy*, *66*, 222–238. https://doi.org/10.1530/EJE-14-0355

Awojobi, O., & Jenkins, G. P. (2016). Managing the cost overrun risks of hydroelectric dams: An application of reference class forecasting techniques. *Renewable and Sustainable Energy Reviews*, *63*, 19–32. https://doi.org/10.1016/j.rser.2016.05.006

Ayub, B., Thaheem, M. J., & Ullah, F. (2019). Contingency release during project execution: The contractor's decision-making dilemma. *Project Management Journal*, *50*(6), 734–748. https://doi.org/10.1177/8756972819848250

Barraza, G. A., Asce, M., & Bueno, R. A. (2007). Cost contingency management. *Journal of Management in Engineering*, *23*(3), 140–146. https://doi.org/10.1061/(ASCE)0742-597X(2007)23:3(140)

Bangladesh Bureau of Statistics. (2020). http://www.bbs.gov.bd/site/page/dc2bc6ce-7080-48b3-9a04-73cec782d0df/-.bbs.gov.bd

Bhargava, A., Labi, S., Chen, S., Saeed, T. U., & Sinha, K. C. (2017). Predicting cost escalation pathways and deviation severities of infrastructure projects using risk-based econometric models and Monte Carlo simulation. *Computer-Aided Civil and Infrastructure Engineering*, *32*(8), 620–640. https://doi.org/10.1111/mice.12279

Bilal, M., & Oyedele, L. O. (2020). Guidelines for applied machine learning in construction industry – A case of profit margins estimation. *Advanced Engineering Informatics*, *43*, 101013. https://doi.org/10.1016/j.aei.2019.101013

Bordat, C., McCullouch, B. G., Labi, S., & Sinha, K. (2004). *An analysis of cost overruns and time delays of INDOT projects* (Publication FHWA/IN/JTRP-2004/07). Joint Transportation Research Program, Indiana Department of Transportation and Purdue University, West Lafayette, Indiana. https://doi.org/10.5703/1288284313134

Chakraborty, D., Elhegazy, H., Elzarka, H., & Gutierrez, L. (2020). A novel construction cost prediction model using hybrid natural and light gradient boosting. *Advanced Engineering Informatics*, *46*, 101201. https://doi.org/10.1016/j.aei.2020.101201

Chang, C. Y., & Ko, J. W. (2017). New approach to estimating the standard deviations of lognormal cost variables in the Monte Carlo analysis of construction risks. *Journal of Construction Engineering and Management*, *143*(1), 06016006. https://doi.org/10.1061/(ASCE)CO.1943-7862.0001207

Curram, S. P., & Mingers, H. (2017). Neural networks, decision tree induction and discriminant analysis: an empirical comparision. *Journal of the Operational Research Society*, *45*(4), 440–450. https://doi.org/10.1057/jors.1987.44

Diab, M. F., Varma, A., & Panthi, K. (2017). Modeling the construction risk ratings to estimate the contingency in highway projects. *Journal of Construction Engineering and Management*, *143*(8), 04017041. https://doi.org/10.1061/(ASCE)CO.1943-7862.0001334

Dursun, O., & Stoy, C. (2016). Conceptual estimation of construction costs using the multistep ahead approach. *Journal Construction Engineering and Management*, *142*(9), 04016038. https://doi.org/10.1061/(ASCE)CO.1943-7862.0001150

Elfahham, Y. (2019). Estimation and prediction of construction cost index using neural networks, time series, and regression. *Alexandria Engineering Journal*, *58*(2), 499–506. https://doi.org/10.1016/j.aej.2019.05.002

Elith, J., Leathwick, J. R., & Hastie, T. (2008). A working guide to boosted regression trees. *Journal of Animal Ecology*, *77*(4), 802–813. https://doi.org/10.1111/j.1365-2656.2008.01390.x

Elmousalami, H. H. (2020a). Artificial intelligence and parametric construction cost estimate modeling: State-of-the-art review. *Journal of Construction Engineering and Management*, *146*(1), 03119008. https://doi.org/10.1061/(ASCE)CO.1943-7862.0001678

Elmousalami, H. H. (2020b). Comparison of artificial intelligence techniques for project conceptual cost prediction: A case study and comparative analysis. *IEEE Transactions on Engineering Management*, *68*(1), 183–196. https://doi.org/10.1109/TEM.2020.2972078

Eybpoosh, M., Dikmen, I., & Birgonul, M. T. (2011). Identification of risk paths in international construction projects using structural equation modeling. *Journal of Construction Engineering and Management*, *137*(12), 1164–1175. https://doi.org/10.1061/(ASCE)CO.1943-7862.0000382

Gazder, U., Islam, M. S., & Arifuzzaman, M. (2020). Parametric modeling of the cost of power plant projects. In *2020 International Conference on Innovation and Intelligence for Informatics, Computing and Technologies* (*3ICT 2020*), Sakheer, Bahrain. https://doi.org/10.1109/3ICT51146.2020.9311963

Gilbert, A., Sovacool, B. K., Johnstone, P., & Stirling, A. (2017). Cost overruns and financial risk in the construction of nuclear power reactors: A critical appraisal. *Energy Policy*, *102*, 644–649. https://doi.org/10.1016/j.enpol.2016.04.001

Gong, H., Sun, Y., Shu, X., & Huang, B. (2018). Use of random forests regression for predicting IRI of asphalt pavements. *Construction and Building Materials*, *189*, 890–897. https://doi.org/10.1016/j.conbuildmat.2018.09.017

Günaydin, H. M., & Doğan, S. Z. (2004). A neural network approach for early cost estimation of structural systems of buildings. *International Journal of Project Management*, *22*(7), 595–602. https://doi.org/10.1016/j.ijproman.2004.04.002

Gunduz, M., & Sahin, H. B. (2015). An early cost estimation model for hydroelectric power plant projects using neural networks and multiple regression analysis. *Journal of Civil Engineering and Management*, *21*(4), 470–477. https://doi.org/10.3846/13923730.2014.890657

Hammad, M. W., Abbasi, A., & Ryan, M. J. (2016). Allocation and management of cost contingency in projects. *Journal of Management in Engineering*, *32*(6), 04016014. https://doi.org/10.1061/(ASCE)ME.1943-5479.0000447

Haque, M. A. (2020). *Bangladesh power sector: An appraisal from a multi-dimensional perspective* (Issue 03 September). https://www.arx.cfa/~/media/AD0129173C34401196A0DA6F7C338035.ashx

Hashemi, S. T., Ebadati, O. M., & Kaur, H. (2019). A hybrid conceptual cost estimating model using ANN and GA for power plant projects. *Neural Computing and Applications*, *31*(7), 2143–2154. https://doi.org/10.1007/s00521-017-3175-5

Hegazy, T., & Ayed, A. (1998). Neural network model for parametric cost estimation of highway projects. *Journal of Construction Engineering and Management*, *124*(3), 210–218. https://doi.org/10.1061/(ASCE)0733-9364(1998)124:3(210)

Hoseini, E., Bosch-Rekveldt, M., & Hertogh, M. (2020). Cost contingency and cost evolvement of construction projects in the preconstruction phase. *Journal of Construction Engineering and Management*, *146*(6), 05020006. https://doi.org/10.1061/(ASCE)CO.1943-7862.0001842

Idrus, A., Nuruddin, M. F., & Rohman, M. A. (2011). Development of project cost contingency estimation model using risk analysis and fuzzy expert system. *Expert Systems with Applications*, *38*(3), 1501–1508. https://doi.org/10.1016/j.eswa.2010.07.061

International Monetary Fund. (2020). *Bangladesh's GDP and inflation rate*. https://www.imf.org/en/Countries/BGD

Islam, M. S., & Nepal, M. (2016). A Fuzzy-Bayesian Model for risk assessment in power plant projects. *Procedia Computer Science*, *100*, 963–970. https://doi.org/10.1016/j.procs.2016.09.259

Islam, M. S., Nepal, M. P., & Skitmore, M. (2018). Modified Fuzzy Group Decision Making Approach to the cost overrun risk assessment of power plant projects. *Journal of Construction Engineering and Management*, *145*(2), 04018126-1–04018126-15. https://doi.org/10.1061/(ASCE)CO.1943-7862.0001593

Islam, M. S., Nepal, M. P., Skitmore, M., & Kabir, G. (2019). A knowledge-based expert system to assess power plant project cost overrun risks. *Expert Systems with Applications*, *136*, 12–32. https://doi.org/10.1016/j.eswa.2019.06.030

Islam, R., Ahmad Bashawir, A. G., Mahyudin, E., & Manickam, N. (2017). Determinants of factors that affecting inflation in Malaysia. *International Journal of Economics and Financial Issues*, *7*(2), 355–364.

Islam, M. S., Nepal, M. P., Skitmore, M., & Drogemuller, R. (2021). Risk induced contingency cost modeling for power plant projects. *Automation in Construction*, *123*, 103519. https://doi.org/10.1016/j.autcon.2020.103519

Jung, J. H., Kim, D. Y., & Lee, H. K. (2016). The computer-based contingency estimation through analysis cost overrun risk of public construction project. *KSCE Journal of Civil Engineering*, *20*(4), 1119–1130. https://doi.org/10.1007/s12205-015-0184-8

Lam, T. Y. M., & Siwingwa, N. (2017). Risk management and contingency sum of construction projects. *Journal of Financial Management of Property and Construction*, *22*(3), 237–251. https://doi.org/10.1108/JFMPC-10-2016-0047

Lee, K. P., Lee, H. S., Park, M., Kim, D. Y., & Jung, M. (2017). Management-reserve estimation for international construction projects based on risk-informed k-NN. *Journal of Management in Engineering*, *33*(4), 04017002. https://doi.org/10.1061/(ASCE)ME.1943-5479.0000510

Lhee, S. C., Issa, R. R. A., & Flood, I. (2012). Prediction of financial contingency for asphalt resurfacing projects using artificial neural networks. *Journal of Construction Engineering and Management*, *138*(1), 22–30. https://doi.org/10.1061/(ASCE)CO.1943-7862.0000408

Li, Y., & Wang, X. (2018). Risk assessment for public–private partnership projects: using a fuzzy analytic hierarchical process method and expert opinion in China. *Journal of Risk Research*, *21*(8), 952–973. https://doi.org/10.1080/13669877.2016.1264451

Loh, W. Y. (2014). Fifty years of classification and regression trees. *International Statistical Review*, *82*(3), 329–348. https://doi.org/10.1111/insr.12016

Maas, C. J., & Hox, J. J. (2005). Sufficient sample sizes for multilevel modeling. *Methodology: European Journal of Research Methods for the Behavioral and Social Sciences*, *1*(3), 86–92. https://doi.org/10.1027/1614-2241.1.3.86

Maronati, G., & Petrovic, B. (2019). Estimating cost uncertainties in nuclear power plant construction through Monte Carlo sampled correlated random variables. *Progress in Nuclear Energy*, *111*, 211–222. https://doi.org/10.1016/j.pnucene.2018.11.011

Mawlana, M., & Hammad, A. (2015). Joint probability for evaluating the schedule and cost of stochastic simulation models. *Advanced Engineering Informatics*, *29*(3), 380–395. https://doi.org/10.1016/j.aei.2015.01.005

Moisen, G. G. (2008). Classification and regression trees. In S. E. Jørgensen, & B. D. Fath (Eds.), *Encyclopedia of ecology* (Vol. 1, pp. 582–588). Elsevier.

Musarat, M. A., Alaloul, W. S., & Liew, M. S. (2021). Impact of inflation rate on construction projects budget: A review. *Ain Shams Engineering Journal*, *12*(1), 407–414. https://doi.org/10.1016/j.asej.2020.04.009

Olaniran, O. J. (2015). The effects of cost-based contractor selection on construction project performance. *Journal of Financial Management of Property and Construction*, *20*(3), 235–251. https://doi.org/10.1108/JFMPC-06-2014-0008

Perner, P., Zscherpel, U., & Jacobsen, C. (2001). A comparison between neural networks and decision trees based on data from industrial radiographic testing. *Pattern Recognition Letters*, *22*(1), 47–54. https://doi.org/10.1016/S0167-8655(00)00098-2

Prasad, A. M., Iverson, L. R., & Liaw, A. (2006). Newer classification and regression tree techniques: Bagging and random forests for ecological prediction. *Ecosystems*, *9*(2), 181–199. https://doi.org/10.1007/s10021-005-0054-1

Razi, M. A., & Athappilly, K. (2005). A comparative predictive analysis of neural networks (NNs), nonlinear regression and classification and regression tree (CART) models. *Expert Systems with Applications*, *29*(1), 65–74. https://doi.org/10.1016/j.eswa.2005.01.006

Salah, A. (2015). *Fuzzy set-based risk management for construction projects*. Concordia University. https://spectrum.library.concordia.ca/980339/

Salah, A., & Moselhi, O. (2015). Contingency modelling for construction projects using fuzzy-set theory. *Engineering, Construction and Architectural Management*, *22*(2), 214–241. https://doi.org/10.1108/ECAM-03-2014-0039

Shaaban, K., & Pande, A. (2016). Classification tree analysis of factors affecting parking choices in Qatar. *Case Studies on Transport Policy*, *4*(2), 88–95. https://doi.org/10.1016/j.cstp.2015.11.002

Shahtaheri, M., Haas, C. T., & Salimi, T. (2016). A stochastic simulation approach for the integration of risk and uncertainty into megaproject cost and schedule estimates. *Construction Research Congress*, *4*, 1669–1679. https://doi.org/10.1061/9780784479827.062

Shahtaheri, M., Haas, C. T., & Rashedi, R. (2017). Applying very large scale integration reliability theory for understanding the impacts of type II risks on megaprojects. *Journal of Management in Engineering*, *33*(4), 04017003. https://doi.org/10.1061/(ASCE)ME.1943-5479.0000504

Singh, A. S., & Masuku, M. B. (2013). Fundamentals of applied research and sampling techniques. *International Journal of Medical and Applied Sciences*, *2*(4), 124–132.

Sonmez, R., Ergin, A., & Birgonul, M. T. (2007). Quantitative methodology for determination of cost contingency in international projects. *Journal of Management in Engineering*, *23*(1), 35–39. https://doi.org/10.1061/(ASCE)0742-597X(2007)23:1(35)

Sovacool, B. K., Gilbert, A., & Nugent, D. (2014). An international comparative assessment of construction cost overruns for electricity infrastructure. *Energy Research & Social Science*, *3*, 152–160. https://doi.org/10.1016/j.erss.2014.07.016

Steinberg, D. (2009). CART: Classification and regression trees. In *The top ten algorithms in data mining* (pp. 193–216). Chapman and Hall/CRC. https://doi.org/10.1201/9781420089653.ch10

Strobl, C., Malley, J., & Tutz, G. (2009). Characteristics of classification and regression trees, bagging and random forests. *Psychological Methods*, *14*(4), 323–348. https://doi.org/10.1037/a0016973

Thal, A. E., Cook, J. J., & Iii, E. D. W. (2010). Estimation of cost contingency for air force construction projects. *Journal of Construction Engineering and Management*, *136*(11), 1181–1188. https://doi.org/10.1061/(ASCE)CO.1943-7862.0000227

Timofeev, R. (2004). *Classification and regression trees* (*CART*) *theory and applications* [Master thesis]. Center of Applied Statistics and Economics, Humboldt University, Berlin.

Touran, A. (2003). Probabilistic model for cost contingency. *Journal of Construction Engineering and Management*, *129*(3), 280–284. https://doi.org/10.1061/(ASCE)0733-9364(2003)129:3(280)

Uzzafer, M. (2013). A contingency estimation model for software projects. *International Journal of Project Management*, *31*(7), 981–993. https://doi.org/10.1016/j.ijproman.2012.12.002

Williams, T. P., & Gong, J. (2014). Predicting construction cost overruns using text mining, numerical data and ensemble classifiers. *Automation in Construction*, *43*, 23–29. https://doi.org/10.1016/j.autcon.2014.02.014

Xia, N., Wang, X., Wang, Y., Yang, Q., & Liu, X. (2017). Lifecycle cost risk analysis for infrastructure projects with modified Bayesian networks. *Journal of Engineering, Design and Technology*, *15*(1), 79–103. https://doi.org/10.1108/JEDT-05-2015-0033

Zhao, Y., Xiang, J., Xu, J., Li, J., & Zhang, N. (2019). Study on the comprehensive benefit evaluation of transnational power networking projects based on multi-project stakeholder perspectives. *Energies*, *12*(2), 249. https://doi.org/10.3390/en12020249

APPENDIX

Table A.1. Data collected from 58 Bangladeshi power plant projects

Project ID	Variables									
	Location	Inflation rate	Project ownership	Plant type	Power (MW)	Contract type	Contingency (%)	Total GDP (USD million)	Construction GDP (USD million)	Project cost (USD million)
1	Rural	2.5	Public	CCPP	90	EPC	8	54590	1814.222	89.86
2	Rural	7	Public	CCPP	100	EPC	8	184010	4718.573	95
3	Urban	7	Public	CCPP	150	EPC	8	184010	4718.573	175
4	Peri-urban	7.5	Public	CCPP	150	EPC	10	161300	4609.806	187.5
5	Peri-urban	11.5	IPP	CCPP	150	EPC	10	131080	3966.103	152.06
6	Rural	11.5	Semi-autonomous	CCPP	150	EPC	10	131080	3966.103	157.5
7	Rural	6.2	Public	CCPP	225	EPC	1	141710	4266.197	250
8	Peri-urban	9.4	Public	CCPP	225	EPC	5	122040	3723.714	200
9	Rural	7	Semi-autonomous	CCPP	225	EPC	10	184010	4718.573	299.42
10	Peri-urban	7.5	Semi-autonomous	CCPP	225	EPC	5	161300	4609.806	200
11	Rural	11.5	Public	CCPP	225	EPC	5	131080	3966.103	450
12	Rural	6.2	Public	CCPP	225	EPC	10	141710	4266.197	312.11
13	Rural	7.5	Public	CCPP	330	EPC	10	161300	4609.806	365.5
14	Peri-urban	6.2	Semi-autonomous	CCPP	335	EPC	8	208320	5124.335	350
15	Peri-urban	7	Semi-autonomous	CCPP	335	EPC	5.59	184010	4718.573	313.62
16	Rural	11.5	IPP	CCPP	350	EPC	8	131080	3966.103	390
17	Peri-urban	5.7	Semi-autonomous	CCPP	350	EPC	8	235620	5563.015	430
18	Peri-urban	7	Semi-autonomous	CCPP	350	EPC	3	184010	4718.573	380.15
19	Rural	7.5	Public	CCPP	365	EPC	5	161300	4609.806	322
20	Peri-urban	6.2	Semi-autonomous	CCPP	400	EPC	8	208320	5124.335	375
21	Rural	5.7	Public	CCPP	400	EPC	8	235620	5563.015	325.65
22	Peri-urban	5.7	Semi-autonomous	CCPP	400	EPC	5	235620	5563.015	500
23	Peri-urban	6.2	Public	CCPP	400	Turnkey	8	208320	5124.335	400
24	Rural	7.5	Public	CCPP	400	EPC	8	161300	4609.806	430.5
25	Peri-urban	7.5	Public	CCPP	400	EPC	3	161300	4609.806	350
26	Rural	9.4	Public	CCPP	410	EPC	8	122040	3723.714	487.65
27	Peri-urban	5.7	Semi-autonomous	CCPP	412	EPC	8	235620	5563.015	508
28	Rural	11.5	Semi-autonomous	CCPP	412	EPC	1	131080	3966.103	328
29	Rural	9.1	Semi-autonomous	CCPP	412	EPC	10	85600	3144.745	355.2

End of Table A1

Project ID	Variables									
	Location	Inflation rate	Project ownership	Plant type	Power (MW)	Contract type	Contingency (%)	Total GDP (USD million)	Construction GDP (USD million)	Project cost (USD million)
30	Rural	6.2	Semi-autonomous	CCPP	420	EPC	10	141710	4266.197	430
31	Peri-urban	5.7	Semi-autonomous	CCPP	450	EPC	10	235620	5563.015	400
32	Rural	3.7	Semi-autonomous	CCPP	450	EPC	10	57500	2140.92	289
33	Rural	5.5	Semi-autonomous	CCPP	718	EPC	10	317470	7291.907	833
34	Rural	7.5	Semi-autonomous	Coal	275	EPC	1	161300	4609.806	337.5
35	Peri-urban	5.5	Public	Coal	400	EPC	10	317470	7291.907	430
36	Rural	5.7	Public	Coal	660	EPC	3	235620	5563.015	450
37	Rural	5.6	Public	Coal	600	EPC	7.2	288420	6651.478	2250
38	Rural	6.2	IPP	HFO	50	BOO	8	141710	4266.197	97.5
39	Rural	11.5	IPP	HFO	50	Turnkey	8	131080	3966.103	49.9
40	Urban	11.5	IPP	HFO	50	BOO	5	131080	3966.103	43.75
41	Peri-urban	7	Public	HFO	100	EPC	8	184010	4718.573	125
42	Rural	7	IPP	HFO	100	BOO	8	184010	4718.573	100
43	Peri-urban	7.5	IPP	HFO	100	EPC	8	161300	4609.806	125
44	Peri-urban	7	IPP	HFO	100	EPC	3	184010	4718.573	104.16
45	Rural	11.5	IPP	HFO	100	Turnkey	8	131080	3966.103	93
46	Peri-urban	11.5	IPP	HFO	100	EPC	8	131080	3966.103	97.5
47	Peri-urban	11.5	IPP	HFO	100	Turnkey	8	131080	3966.103	110
48	Peri-urban	5.6	Semi-autonomous	HFO	100	EPC	3	288420	6651.478	95
49	Peri-urban	7.5	Public	HFO	100	Turnkey	10	161300	4609.806	142.7
50	Peri-urban	11.5	Public	Natural gas	50	EPC	8	131080	3966.103	65
51	Rural	7.5	IPP	Natural gas	25	BOO	8	161300	4609.806	35
52	Peri-urban	6.2	Semi-autonomous	Natural gas	50	BOO	8	141710	4266.197	75
53	Rural	7	Public	Natural gas	60	EPC	8	184010	4718.573	90
54	Rural	3.7	IPP	Natural gas	80	BOO	8	57500	2140.92	40
55	Rural	7.5	IPP	Natural gas	100	BOO	8	161300	4609.806	150
56	Urban	2.5	Public	Natural gas	150	EPC	3	54590	1814.222	210.4
57	Rural	3	Public	Natural gas	420	Turnkey	8	38230	993.6735	189
58	Rural	10.1	Public	Natural gas	400	Turnkey	8	45920	1189.671	167