Modelos para estimativa de custos com o uso de regressão linear: modelagem com obras penitenciárias

Models for early cost estimating using linear regression: penitentiary projects modeling

Leandro Modesto Prates Beltrão
Michele Tereza Marques Carvalho
Raquel Naves Blumenschein
Álvaro Teixeira de Paiva
Maíra Vitoriano Rodrigues de Freitas

Resumo

A estimativa de custos de empreendimentos da construção civil em fases preliminares do projeto nem sempre é tarefa fácil, envolvendo elevado grau de imprecisão e incerteza. Ainda que em estágios iniciais, inconsistências nessas estimativas têm o potencial de gerar prejuízos financeiros e inviabilizar a realização de obras. Diante disso, o objetivo central desta pesquisa é apresentar uma estrutura para o desenvolvimento de modelos destinados à estimativa de custos, por meio da técnica de regressão linear. O método proposto para a estrutura divide-se em cinco fases: (1) definição de requisitos do modelo, (2) seleção de variáveis independentes, (3) coleta e elaboração do banco de dados, (4) modelagem dos dados e (5) avaliação da performance do modelo. A estrutura foi testada em obras de construção de cadeias e penitenciárias federais, visando comprovar sua aplicabilidade. A aplicação da regressão retornou dois modelos válidos, com margens de erro de 23% e 25%. A estrutura em si representa uma das principais contribuições desta pesquisa, podendo ser replicada por diversos agentes na construção dos próprios modelos para a estimativa de custos.

Palavras-chave: Construção civil. Estimativa de custos. Regressão linear.

[1]Leandro Modesto Prates Beltrão
[1]Universidade de Brasília
Brasília – DF – Brasil

[2]Michele Tereza Marques Carvalho
[2]Universidade de Brasília
Brasília – DF – Brasil

[3]Raquel Naves Blumenschein
[3]Universidade de Brasília
Brasília – DF – Brasil

[4]Álvaro Teixeira de Paiva
[4]Fundação Getulio Vargas
Brasília – DF – Brasil

[5]Maíra Vitoriano Rodrigues de Freitas
[5]Empresa de Planejamento e Logística
Brasília – DF – Brasil

Recebido em 21/12/21
Aceito em 15/01/22

Abstract

Early cost estimation of construction projects is not an easy task, as such projects involve a high level of inaccuracy and uncertainty. Even in the early stages, errors in the estimates can result in financial loss and jeopardize construction completion. Therefore, the main objective of this research study is to present a framework for building cost estimation models, using the linear regression technique. The framework method is divided into five phases: (1) identifying the model's requirements, (2) selection of the independent variables, (3) database construction, (4) data modelling and (5) model performance evaluation. A case study was conducted on federal penitentiary construction projects to test the applicability of the framework. Through the case study, two valid models were built, and their margins of error were 23 and 25%. The framework itself is one of the main contributions of this study, and it can be replicated by practitioners to develop models for construction cost estimation.

***Keywords**: Construction. Cost estimation. Linear regression.*

Introdução

No cenário da construção civil, a estimativa de custos em fases preliminares de desenvolvimento do projeto, como estudo de viabilidade técnico-econômica ou mesmo anteprojeto de engenharia, é reconhecidamente uma tarefa de difícil execução, uma vez que as informações disponíveis possuem pouco detalhamento e considerável nível de imprecisão. Em consequência, referenciais de custo confiáveis e passíveis de aplicação em tais fases são instrumentos primordiais à programação de investimentos e ao planejamento estratégico, além de auxiliarem a elaboração e a avaliação de orçamentos.

Atualmente no Brasil há dois principais indicadores voltados à estimativa de custos de obras de edificações. São eles:

(a) custo unitário básico de construção (CUB): indicador de custos de obras de construção civil calculado e divulgado pelos Sindicatos da Indústria da Construção Civil (Sinduscons) estaduais; e

(b) custo médio do metro quadrado na construção civil: indicador de custos de obras de construção civil produzido conjuntamente pelo Instituto Brasileiro de Geografia e Estatística (IBGE) e pela Caixa Econômica Federal.

Em que pese a relevância nacional de ambos os indicadores, eles não são apropriados para retratar a formação de custos de tipologias construtivas com alto grau de especificidade, como é o caso de edificações penais do tipo cadeia pública e penitenciária. Diante disso, o desenvolvimento de referenciais de custos bem fundamentados e confiáveis para estimativas em fases iniciais de maturidade do projeto ganha relevância no âmbito da construção civil. Tais referenciais têm o potencial de auxiliar a gestão de custos como um todo, pois são fontes mais precisas para embasar a tomada de decisão sobre custo-benefício, permitem mais agilidade na previsão orçamentária, tendem a reduzir riscos de inviabilização do empreendimento por falhas na etapa de planejamento, dentre outros benefícios. Além disso, referenciais precisos e consistentes para a estimativa de custos podem inclusive promover economia e contribuir para projetos mais sustentáveis (RAFIEI; ADELI, 2018).

Em se tratando de níveis de precisão de estimativa de custos e de orçamento, a Prática Recomendada nº 17R-97 (ASSOCIATION..., 2011) da *Association for the Advancement of Cost Engineering* (AACE) fornece diretrizes de uso consolidado na construção civil, conforme disposto na Tabela 1.

Em analogia aos níveis de maturidade do projeto da Tabela 1, é possível assumir que a classe 5 corresponde ao programa de necessidades; a 4, ao estudo de viabilidade técnico-econômica; a 3, ao anteprojeto; a 2, ao projeto básico; e a 1, ao projeto executivo.

Observa-se que a Prática Recomendada n nº 17R-97 (ASSOCIATION..., 2011) recomenda metodologias estocásticas para a estimativa de custos das classes 4 e 5, a qual deve apresentar níveis de precisão de 40 a 85% e de 0 a 80%, respectivamente. Em suma, metodologias estocásticas envolvem a modelagem entre custos e características do objeto em análise, sendo baseadas geralmente em dados históricos de objetos similares realizados no passado. Dentre os diversos métodos estocásticos existentes, destaca-se, no âmbito deste estudo, a regressão linear.

Tabela 1 – **Níveis de precisão segundo níveis de maturidade do projeto**

Classe	Nível de maturidade do projeto	Metodologia	Margens de erroI	Níveis de precisão
5	0 a 2%	Estocástica ou julgamento	± 20 a ± 100%	0 a 80%
4	1 a 15%	Principalmente estocástica	± 15 a ± 60%	40 a 85%
3	10 a 40%	Mista, mas principalmente estocástica	± 10 a ± 30%	70 a 90%
2	30 a 75%	Principalmente determinística	± 5 a ± 15%	85 a 95%
1	65 a 100%	Determinística	± 5	95%

Fonte: adaptada de AACE International (ASSOCIATION..., 2011),

Ante o exposto, o objetivo deste artigo é apresentar uma estrutura para o desenvolvimento de modelos de regressão linear voltados à estimativa de custos, a qual seja passível de ser facilmente replicada por instituições públicas e privadas. A estrutura deve abordar todas as fases e processos necessários à construção do modelo, além de indicar ferramentas que possibilitem sua implementação. Outro objetivo desta pesquisa é apresentar o indicador "fator de influência da variável", concebido para sistematizar a seleção das variáveis que devem compor a modelagem. A estrutura proposta é aplicada em obras de construção de cadeias públicas e penitenciárias do Departamento Penitenciário Nacional (DEPEN) do Ministério da Justiça e Segurança Pública (MJSP) para testar sua validade.

Referencial teórico

O referencial teórico que embasou esta pesquisa encontra-se sintetizado em dois tópicos:

(a) modelos para a estimativa de custos; e

(b) regressão linear: principais conceitos e pressupostos.

Modelos para a estimativa de custos

Na literatura pertinente à construção civil, nota-se o emprego crescente da modelagem com base em dados históricos para a estimativa de custos em fases preliminares de desenvolvimento do projeto. Sanders, Maxwell e Glagola (1992) apresentaram um modelo para estimativa de custos de obras de alargamento de pontes, o qual foi construído por meio da técnica estatística de regressão e testado pelo processo *leave-one-out*, chegando à margem de erro de 6%.

Hegazy e Ayed (1998) fizeram uso da técnica de redes neurais para a estimativa de custos de obras rodoviárias. Para tanto, os autores utilizaram uma base de dados composta por 18 projetos e destacaram 10 fatores principais capazes de descrever uma rodovia e formar seu custo. Para o cálculo dos pesos da rede neural, os autores testaram os métodos (1) *backpropagation*, (2) otimização simplex e (3) algoritmo genético, obtendo o menor erro (20%) para o método (2).

Moselhi e Siqueira (1998) apresentaram um sistema automático para previsão de custos de estruturas metálicas (*steel framing*), em que foram analisados 34 projetos de uma fábrica no Canadá. Os autores utilizaram cerca de 15% da amostra para validar seus resultados, chegando a erros de 11% e 15% para redes neurais e regressão, respectivamente. Já nos Estados Unidos, Williams (2003) utilizou regressões logarítmicas para estimar os custos de obras rodoviárias com elevada precisão a partir do tratamento de dados históricos de cinco agências governamentais.

O estudo de Kim, An e Kang (2004) comparou a acurácia de três técnicas para estimativa de custos de obras: regressão linear múltipla, redes neurais e o *Case-Based Reasoning* (CBR). Para uma amostra de 530 projetos de obras residenciais, o melhor modelo obtido pelas redes neurais forneceu resultados mais apurados dentre as três técnicas, apesar de apresentar desvantagens em termos do tempo para aquisição dos dados e para modelagem por tentativa e erro.

Lowe, Emsley e Harding (2006) descreveram o desenvolvimento de modelos de regressão linear múltipla a partir de uma base de dados de 286 projetos de construção situados no Reino Unido. Lançando mão de variáveis independentes conhecidas em etapas iniciais de projeto, a regressão linear obteve resultados favoráveis quando comparados aos métodos tradicionais de estimativa de custos, com margens de erro na faixa de 25%.

Por meio da regressão linear múltipla, Petroutsatou, Lambropoulos e Pantouvakis (2006) desenvolveram modelos para estimativa de custos de túneis para rodovias, a partir de uma amostra de 33 túneis gregos (totalizando 46 quilômetros de extensão). Correlacionando parâmetros geotécnicos com os custos, os autores chegaram a um modelo com precisão de cerca de 88%. Já Rostami *et al*. (2013) elaboraram, com sucesso, modelos de estimativa de custo de túneis a partir de dados de 272 projetos desenvolvidos nos Estados Unidos, baseando-se na técnica de regressão múltipla.

Na Turquia, Sonmez (2008) propôs uma metodologia conjunta entre a análise de regressão e o método de reamostragem *bootstrap*, para o desenvolvimento de estimativas de custos a partir de uma base de dados de 20 projetos de construção. A combinação das técnicas de regressão e *bootstrap* retornou uma margem de erro de 12% para o melhor modelo. Por sua vez, Sonmez e Ontepeli (2009) apresentaram um modelo que combina redes neurais com regressões após analisarem dados de construção de 13 linhas de metrô.

Já Kim (2011) coletou dados históricos de 123 pontes ferroviárias para propor uma metodologia de estimativa de custos baseada em algoritmos de regressão múltipla, a qual se mostrou 30% mais eficaz do que os modelos até então utilizados pelo Ministério de Construção e Transporte da Coreia do Sul. Kim e Hong (2012) revisaram a proposta por Kim (2011) por meio de análises de regressão, reduzindo o erro da metodologia original em 16%.

Mahamid (2011) elaborou um modelo de regressão linear múltipla para estimativa de custos iniciais envolvidos na construção de rodovias. Como resultado, o estudo identificou que variáveis independentes que representam quantidades de projeto geram modelos de melhor desempenho, quando comparadas com variáveis que caracterizam a rodovia (e.g. região geográfica, relevo etc.)

Asmar, Hanna e Whited (2011) conceberam uma metodologia similar ao *Program Evaluation and Review Technique* (PERT) para estimar custos de obras rodoviárias ainda na fase conceitual, onde somente 30% do projeto está completo. A partir de uma base de dados constituída de 77 obras do Departamento de Transporte (DOT) de Wisconsin, os autores chegaram a um erro de 20% para seu modelo.

Zhu, Feng e Zhou (2010) e Zhai, Jiang e Pedrycz (2012) se valeram da lógica *fuzzy* para investigar modelos de predição de custo na construção civil. Enquanto os primeiros autores combinaram a teoria dos conjuntos *fuzzy* com redes neurais, os segundos propuseram aprimoramentos ao algoritmo *fuzzy c-means* (FCM) para a clusterização (i.e., divisão da amostra em grupos).

Tanto Hollar *et al*. (2013) como Liu *et al*. (2013) estudaram a utilização da regressão linear múltipla para estimativa de serviços de engenharia consultiva de obras de infraestrutura de transportes. Em seus estudos, os autores demonstraram a eficácia e a validade da técnica de regressão na previsão de tais custos.

Zhang, Minchin Junior e Agdas (2017) propuseram o processo do *Least Absolute Shrinkage and Selection Operator* (LASSO) como modelo de regressão para prever custos de recapeamento de rodovias. Para uma base de dados de 741 projetos do DOT da Flórida, o LASSO apresentou resultados melhores quando comparado com a regressão gerada pelo método de mínimos quadrados ordinários (MQO).

Petruseva *et al*. (2017) coletaram dados da construção de 75 edificações na Bósnia para comparar os métodos de regressão linear e S*upport Vector Machine* (SVM). Ao final, o modelo construído com o SVM obteve melhor desempenho, com margem de erro inferior a 1%, enquanto o modelo construído com a regressão linear apresentou erro de cerca de 4%.

Ogungbile, Rasak e Oke (2018) estudaram a estimativa de custos de construção de rodovias na Nigéria por meio da regressão linear. Para uma amostra de 97 projetos, os autores chegaram à margem de erro de 8%. Mahalakshmi e Rajasekaran (2019) também investigaram os custos de construção de rodovias, mas pela técnica de redes neurais, obtendo modelos com margem de erro da ordem de 8%.

Elmousalami, Elyamany e Ibrahim (2018) compararam modelos construídos por regressão linear e redes neurais na previsão de custos de obras de canais para irrigação. Os autores selecionaram inicialmente 17 variáveis independentes para a modelagem, e a amostra do estudo foi composta por 111 projetos. O modelo de melhor performance foi desenvolvido por meio da regressão linear com transformação quadrática na variável dependente, obtendo um erro de aproximadamente 8%.

Nas 36 pesquisas investigadas por Elmousalami (2020) em sua revisão da literatura acerca da modelagem com base em dados históricos para a estimativa de custos na construção civil, modelos híbridos (27%) foram os mais utilizados, seguidos de modelos elaborados por meio de redes neurais (25%), teoria dos conjuntos *fuzzy* (14%), regressão (13%), SVM (11%) e CBR (10%).

Por sua vez, Hashemi, Ebadati e Kaur (2020) conduziram uma revisão sistemática da literatura com objetivos similares aos de Elmousalami (2020). Os autores investigaram cerca de 90 artigos publicados nos últimos 30 anos sobre a estimativa de custos de projetos de construção por meio de modelagem baseada em dados. O estudo indicou que as redes neurais (44%) foi a técnica mais frequente, acompanhada pela regressão linear (21%) e CBR (9%).

Em que pese a quantidade significativa de pesquisas internacionais verificadas, não foram identificados artigos acerca da temática no cenário brasileiro. A ausência de discussões nacionais que evidenciem benefícios e possibilidades de modelos baseados em dados nas mais diversas áreas da construção civil tende a protelar a adequação da cadeia produtiva da construção aos preceitos da Indústria 4.0, a qual propõe como um de seus pilares o uso inteligente de dados na orientação e, sobretudo, integração de processos. Consequentemente, publicações em nível nacional a respeito do tema abordado nesta pesquisa são primordiais para fomentar o desenvolvimento científico.

Ademais, os estudos citados ao longo deste artigo focaram principalmente na conceituação teórica das técnicas, sem aprofundar nas demais fases envolvidas em suas modelagens. A estrutura exibida nesta pesquisa visa preencher essa lacuna, avançando no conhecimento científico ao propor um método detalhado que possibilita o uso da regressão linear para fins de estimativa de custos. Além da estrutura em si, o fator de influência da variável proposto representa outra contribuição relevante deste artigo, possibilitando maior sistematização ao processo de seleção das variáveis independentes mais relevantes ao modelo, etapa crucial para o sucesso da modelagem.

Regressão linear: principais conceitos e pressupostos

Apesar do grande número de estudos e pesquisas sobre a modelagem baseada em dados para a estimativa de custos de obras, Gardner, Gransberg e Jeong (2016) apontam que não há entendimento quanto à técnica de melhor desempenho. Esse apontamento ficou comprovado pela investigação da literatura elucidada no tópico anterior. Ora os autores indicaram determinada técnica como de melhor desempenho, ora outra.

Além do mais, a estrutura proposta neste artigo visa ser utilizada por agentes de instituições públicas e privadas na estimativa de custos; logo, os modelos devem priorizar a simplicidade de aplicação e manuseio, bem como ser de fácil inspeção e revisão periódica.

Posto isso, a presente pesquisa adota a técnica de regressão linear por duas principais razões:

(a) os modelos gerados pela regressão são constituídos majoritariamente por equações lineares de fácil resolução, assim seu emprego é simples e, uma vez construídos, não requerem conhecimentos específicos para sua correta utilização; e

(b) os processos da regressão linear são mais transparentes e auditáveis quando comparados, por exemplo, aos de redes neurais, possibilitando melhor explicação e entendimento dos modelos desenvolvidos.

A conceituação apresentada a seguir sobre a regressão linear fundamenta-se no livro-referência de Montgomery, Peck e Vining (2012).

A Equação 1 exibe a função geral que correlaciona a variável dependente "y", ou variável resposta, com as "k" variáveis independentes "x", ou regressores.

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k + \Delta \quad \text{Eq. 1}$$

Onde "β_0" e "β_j, j = 1, ..., k" são os coeficientes da regressão e "Δ" é o termo de erro do modelo. A Equação 1 retrata um hiperplano de k-ésima dimensões em relação aos regressores "x_j, j = 1, ..., k". Desse modo, a regressão linear pode ser de dois tipos: simples (somente uma variável independente) ou múltipla (duas ou mais variáveis independentes).

O coeficiente "β_j" representa a variação esperada em "y" com a variação unitária em determinado regressor "x_j", quando todos os outros regressores são mantidos constantes. Já o parâmetro "β_0" representa o intercepto da equação. Assim, não possui variável associada a ele.

Na regressão, a força da correlação linear entre as variáveis dependente e independentes é comumente avaliada pelo coeficiente de determinação "R^2". O coeficiente em questão representa a proporção da variabilidade da variável resposta explicada pelas variáveis independentes. Tem-se que "$0 \leq R^2 \leq 1$"; logo, valores de "R^2" próximos a 1 indicam que a maior parte da variabilidade em "y" é explicada por "x_j".

O método dos mínimos quadrados ordinários é largamente empregado no cálculo dos coeficientes da regressão. O MQO visa encontrar o ajuste mais adequado para um conjunto de dados, de modo a minimizar a soma dos quadrados dos erros, ou seja, dos quadrados das diferenças entre os valores observados "y_i" e os valores ajustados "$\hat{y}_i$", sendo "i = 1, 2, ..., n" e "n" o número de observações da amostra. O erro, ou resíduo, "e_i" pode ser computado segundo a Equação 2.

$$e_i = y_i - \hat{y}_i, i = 1, 2, \ldots, n \quad \text{Eq. 2}$$

A estimativa dos coeficientes da regressão pelo MQO assume quatro principais hipóteses:

(a) I: a relação entre a variável dependente e as variáveis independentes é linear;

(b) II: as variáveis independentes apresentam pouca ou nenhuma correlação entre si;

(c) III: os resíduos são normalmente distribuídos, com média igual a zero e variância constante; e

(d) IV: os resíduos são não correlacionados.

A hipótese I está associada à significância estatística do modelo de regressão, ou seja, à existência de relação linear entre as variáveis dependente e independentes. Já a hipótese II tem o intuito de eliminar a multicolinearidade do modelo. Multicolinearidade é o fenômeno que ocorre quando há relação linear forte entre as variáveis independentes, prejudicando consideravelmente a precisão com que os coeficientes são calculados pelo MQO por dois motivos:

(a) a multicolinearidade provoca grandes variâncias e covariâncias para os estimadores dos coeficientes da regressão. Logo, diferentes amostras, coletadas em mesmos níveis de "x", podem levar a coeficientes bastante distintos; e

(b) a multicolinearidade tende a produzir coeficientes inflacionados em valor absoluto.

O pressuposto de normalidade dos resíduos indicado na hipótese III faz-se necessário tanto para que testes de hipóteses sejam válidos quanto para resguardar a confiabilidade do modelo. A hipótese III estabelece ainda que a variância dos resíduos deve ser constante, fato intitulado de homocedasticidade dos resíduos. A suposição de homocedasticidade tem o propósito de mitigar que a variabilidade das variáveis independentes afete a variabilidade dos resíduos. Modelos não homocedásticos, ou seja, heterocedásticos, são afetados de duas principais maneiras:

(a) os erros padrões dos coeficientes não são corretos, logo a inferência estatística não é válida; e

(b) não se pode assegurar que os estimadores do MQO são os que geram mínima variância para os coeficientes da regressão.

A hipótese IV é aplicável apenas em situações em que a evolução temporal dos dados é importante para a modelagem, por isso não é elucidada em detalhes neste artigo.

Por fim, é importante destacar que, dentre as pesquisas trazidas no tópico anterior que adotaram a regressão linear como algoritmo, não foram identificadas análises acerca das hipóteses de regressão elucidadas. As discussões e análises dos resultados focaram principalmente em medidas de avaliação global da performance dos modelos, tais como coeficiente de determinação "R^2" e erro médio percentual absoluto. De maneira geral, a avaliação do atendimento às hipóteses da regressão para validação dos modelos elaborados é negligenciada em artigos científicos das mais variadas áreas. Em consequência, outra contribuição deste trabalho é justamente enfatizar a importância de verificar os pressupostos da regressão linear para que os modelos construídos sejam válidos para a estimativa de custos, apresentando técnicas para a averiguação de cada hipótese.

Método de pesquisa

O método da estrutura para o desenvolvimento de modelos de regressão linear voltados à estimativa de custos em estágios preliminares do projeto divide-se em cinco fases:

(a) definição de requisitos do modelo;

(b) seleção de variáveis independentes;

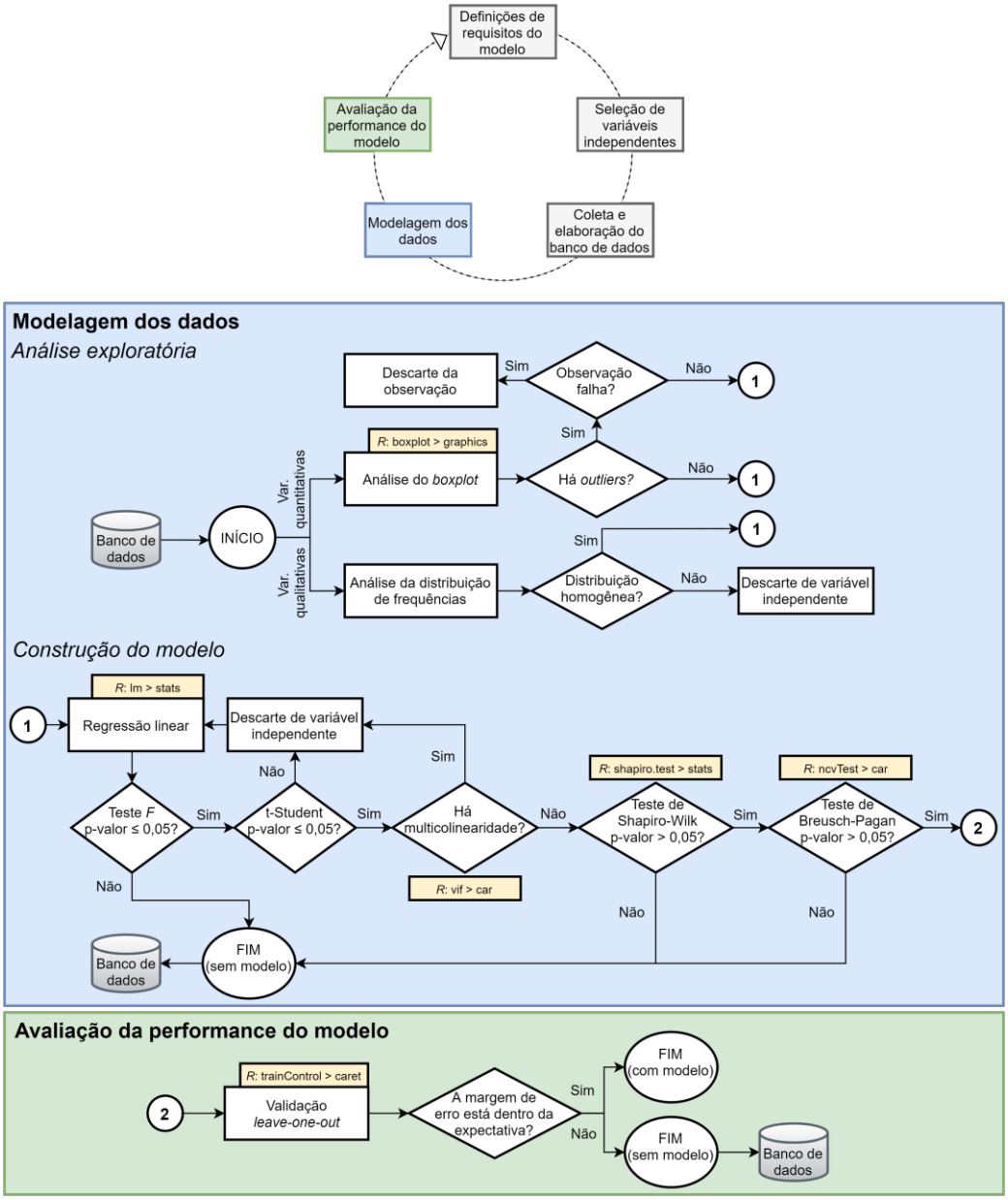
(c) coleta e elaboração do banco de dados;

(d) modelagem dos dados; e

(e) avaliação da performance do modelo.

A Figura 1 apresenta a estrutura em si, com destaque para o detalhamento dos fluxogramas das fases de modelagem dos dados e de avaliação da performance do modelo. O caráter cíclico ilustrado para o relacionamento entre as seis fases almeja indicar que melhorias contínuas devem sempre ser implementadas no modelo, seja com a inclusão de novos dados, seja com o aprendizado resultante de sua aplicação.

Figura 1 – **Estrutura para desenvolvimento de modelos de regressão linear para estimativa de custos**



Definição de requisitos do modelo

Nesta fase inicial, o escopo de aplicação do modelo deve ser suficientemente caracterizado e delimitado. Para tanto, é importante analisar necessidades e expectativas para o modelo, além de obter informações sobre os ambientes interno e externo circundantes. Sugere-se como informações mínimas a serem levantadas:

(a) contexto de aplicação;

(b) contexto em que está inserido;

(c) fase de projeto a que se destina;

(d) finalidade de uso;

(e) futuros usuários;

(f) nível de precisão requerido; e

(g) partes interessadas.

O objetivo desta fase é delimitar diretrizes para nortear todo o desenvolvimento do modelo, otimizando, então, as chances de ele atender ao propósito para o qual foi concebido originalmente.

Seleção de variáveis independentes

Observa-se na literatura que a seleção de variáveis independentes pode ser realizada de diversas maneiras, sendo indispensável experiência, bom senso técnico e conhecimento na tipologia de obra cujos custos estão sendo modelados. Com o intuito de sistematizar a seleção das variáveis independentes mais relevantes ao modelo, propõe-se o seguinte indicador: fator de influência da variável (FIV).

Inicialmente, o cálculo do FIV requer o levantamento de variáveis com reconhecido impacto nos custos das obras. Para tanto, pode-se valer da própria experiência profissional, mas também de consultas com especialistas, análise de projetos passados, investigação da literatura, dentre outras fontes de informação. Em adição, recomenda-se fortemente o estudo das curvas ABC do espaço amostral disponível para a modelagem, buscando identificar variáveis capazes de representar os serviços que compõem a faixa A da curva.

Após o levantamento prévio de variáveis, deve-se proceder com a aplicação de questionários por profissionais com experiência na formação de custos da tipologia de obra modelada. Os questionários devem ser elaborados para que os profissionais possam julgar subjetivamente cada uma das variáveis de modo prático e intuitivo. Por meio deles, os especialistas devem avaliar quatro proposições para cada variável previamente levantada. São elas:

(a) proposição 1: a variável é relevante para os custos da tipologia de obra em questão;

(b) proposição 2: a precisão da variável é satisfatória para a fase de projeto a que se destina o modelo;

(c) proposição 3: há dados históricos em quantidade adequada sobre a variável nos projetos passados disponíveis no espaço amostral; e

(d) proposição 4: os dados de entrada da métrica proposta para a variável podem ser facilmente obtidos na fase de projeto a que se destina o modelo.

A proposição 1 visa avaliar a influência das variáveis nos custos da obra. Já a proposição 2 tem o intuito de analisar o grau de definição da variável na fase de projeto de aplicação do modelo, pois variáveis passíveis de alterações significativas ao longo do detalhamento do projeto podem produzir erros na estimativa. A proposição 3, por sua vez, almeja evitar a seleção de variáveis para as quais não há informações históricas suficientes para o desenvolvimento de modelos de regressão linear, prevenindo a tarefa improdutiva de coletar dados para tais variáveis. Finalmente, pretende-se com a proposição 4 investigar a facilidade de coletar dados da variável na fase de projeto a que se destina o modelo, uma vez que o modelo deve propiciar agilidade à orçamentação. A última proposição fundamentou-se nas conclusões de Gardner, Gransberg e Jeong (2016) e Mahamid (2011), os quais sugeriram variáveis que requerem pouco esforço na obtenção de dados. A avaliação de cada proposição deve ser baseada nas respostas subjetivas indicadas no Quadro 1.

A partir das respostas dispostas no Quadro 1, os profissionais selecionados devem avaliar o nível de concordância com as afirmações trazidas nas proposições, dentro de uma escala que varia de "Concordo fortemente" a "Discordo fortemente", segundo as explicações fornecidas.

A Figura 2 apresenta um exemplo de questionário elaborado para a avaliação das variáveis, cujo intuito é ilustrar um modo simples de aplicar as diretrizes delimitadas previamente neste tópico. Cabe adicionar um campo para que os profissionais possam propor variáveis não levantadas previamente.

Uma vez que as respostas se baseiam em termos qualitativos, elas precisam ser convertidas em termos quantitativos para que o FIV de cada variável possa ser calculado. Para tanto, deve-se utilizar a escala exibida na Tabela 2.

Quadro 1 – **Respostas e suas explicações**

Resposta	Explicação da resposta
Concordo fortemente	A proposição SEMPRE é válida em obras dessa tipologia
Concordo	A proposição é válida NA MAIORIA das obras dessa tipologia
Neutro	A proposição é MODERADAMENTE válida em obras dessa tipologia
Discordo	A proposição NÃO é válida NA MAIORIA das obras dessa tipologia
Discordo fortemente	A proposição NUNCA é válida em obras dessa tipologia

Figura 2 – **Exemplo de questionário para seleção de variáveis independentes**

ID	Variável	Métrica	Proposição 1	Proposição 2	Proposição 3	Proposição 4
1	Variável independente X	Dado de entrada da variável X				
2	Variável independente Y	Dado de entrada da variável Y				
3	Variável independente Z	Dado de entrada da variável Z				
4	Variável independente N	Dado de entrada da variável N				

Proposição
1 - A variável é relevante para os custos da tipologia de obra em questão
2 - A precisão da variável é satisfatória para a fase de projeto a que se destina o modelo
3 - Há dados históricos em quantidade adequada sobre a variável nos projetos passados disponíveis no espaço amostral
4 - Os dados de entrada da métrica proposta para a variável podem ser facilmente obtidos na fase de projeto a que se destina o modelo

Resposta	Explicação da resposta
Concordo fortemente	A proposição **SEMPRE** é válida em obras dessa tipologia
Concordo	A proposição é válida **NA MAIORIA** das obras dessa tipologia
Neutro	A proposição é **MODERADAMENTE** válida em obras dessa tipologia
Discordo	A proposição **NÃO** é válida **NA MAIORIA** das obras dessa tipologia
Discordo fortemente	A proposição **NUNCA** é válida em obras dessa tipologia

Tabela 2 – **Escala de conversão**

Termo qualitativo (resposta)	Termo quantitativo
Concordo fortemente	4
Concordo	3
Neutro	2
Discordo	1
Discordo fortemente	0

Após a conversão das respostas de termo qualitativo para quantitativo, a pontuação "s" de cada variável "j" pode ser calculada por meio da soma dos termos quantitativos "T" de cada proposição "q", para cada especialista "d", conforme a Equação 3. Atribui-se peso 3 a proposição 1 "$T_{q\,=\,1}$", pois ela se refere diretamente à relevância da variável para a formação de custos.

$$s_j{}^d = 3 \times T_1 + \sum_{q=2}^{4} T_q \qquad \text{Eq. 3}$$

Finalmente, o cálculo do FIV de determinada variável independente é efetuado por meio da média aritmética simples entre as pontuações "s" de cada especialista "d", segundo a Equação 4.

$$FIV_j = \frac{\sum_{d=1}^{D} s_j{}^d}{D} \qquad \text{Eq. 4}$$

Onde "D" é o número de respondentes do questionário.

A partir do fator de influência calculado para cada variável, deve-se elaborar um *ranking* de prioridade entre elas. Sugere-se que metade das variáveis de maior classificação sejam selecionadas para o decorrer da modelagem. Essa métrica de seleção baseia-se na curva ABC, segundo a qual os itens de maior relevância (faixa A) constituem aproximadamente 20% da amostra e os itens de relevância intermediária (faixa B) correspondem em torno de 30%, totalizando, assim, 50% da amostra.

Coleta e elaboração do banco de dados

Independentemente da análise almejada (e.g. estimativa de custo, de prazo, de produtividade), modelos de regressão linear sempre requerem dados confiáveis e em quantidade suficiente para apresentarem boa performance. De fato, dados históricos são a fundação da regressão linear, sendo sua qualidade fator chave para a solidez dos resultados da estimativa.

Quanto à coleta de dados referentes aos custos, International Society of Parametric Analysts (2008) apontou que a principal e mais confiável fonte são os registros contábeis da organização para a qual o modelo é desenvolvido. Em adição a esses registros, destacam-se outras fontes de coleta de dados de custos:

(a) contratos;

(b) medições;

(c) orçamentos em nível de projeto executivo;

(d) ordens de compra; e

(e) propostas de licitação.

A lista anterior não visa ser exaustiva, cabendo a cada instituição definir as origens de dados mais confiáveis e representativas da realidade.

Em relação às variáveis independentes, a principal fonte para coleta de dados consiste em toda a documentação do projeto executivo da obra, além das fontes de custos supracitadas.

À medida que os dados forem coletados, eles devem ser armazenados em uma planilha editável, buscando estruturá-los no formato de banco de dados. Dessa forma, cada coluna do banco deve conter os dados de apenas uma variável, seja ela dependente ou independente, ao passo que cada linha deve contemplar os dados de somente uma observação do espaço amostral.

Finalizada a coleta, os dados devem ser ajustados com vistas a tornar o banco consistente e homogêneo. Enquanto alguns ajustes são usuais, como a correção da inflação, outros demandam análises mais aprofundadas, variando caso a caso. Na sequência, exibem-se ajustes aos quais se deve atentar:

(a) anomalias (e.g. greve de trabalhadores, catástrofes naturais, práticas não convencionais de contratação ou fiscalização);

(b) dados com formatos errados ou diferentes;

(c) dados com métricas erradas ou diferentes;

(d) inflação;

(e) mudanças de legislação ou normas; e

(f) mudanças no processo de orçamentação.

Novamente, a lista fornecida não abrange todos os ajustes possíveis, os quais devem ser avaliados dentro do contexto interno e externo em que o modelo está inserido. Cabe destacar que, em se tratando da construção de modelos de regressão linear, é essencial que todos os custos estejam em um mesmo mês-base. Posto isso, é indispensável a correção dos custos quanto à inflação, seja por meio de índices de reajustamento, seja pela própria atualização dos preços que compõem os orçamentos.

Modelagem dos dados

A modelagem dos dados divide-se em dois processos:

(a) análise exploratória; e

(b) construção do modelo.

Adota-se o *software* R Studio como referência para esta metodologia, dado seu caráter livre e integrado para a linguagem de programação R.

Análise exploratória

A análise exploratória visa promover o entendimento inicial sobre o comportamento dos dados da amostra. A análise proposta é do tipo univariada, ou seja, os dados de cada variável (dependente e independentes) devem ser tratados de modo individual.

Para a análise exploratória das variáveis quantitativas contínuas, deve-se utilizar a ferramenta gráfica conhecida como *boxplot*, ou gráfico de caixa. Os *boxplots* podem ser construídos pela função "boxplot" do pacote "graphics" do R.

O *boxplot* é formado por cinco elementos principais: primeiro quartil (25% das menores observações), mediana (50% das menores observações), terceiro quartil (75% das menores observações) e limites inferior e superior (definidos a partir dos valores dos quartis e da amplitude interquartil). Quaisquer valores além dos limites inferior e superior são considerados *outliers*, ou seja, valores extremos e discrepantes do restante da amostra.

Vale destacar que a identificação de *outliers* não deve ser utilizada para a rejeição automática de dados. Análises técnicas também devem embasar a decisão de descartar ou não determinada observação discrepante.

A análise exploratória das variáveis qualitativas deve ser realizada por meio da tabela de distribuição de frequências, avaliando-se as frequências absoluta e percentual dos dados. Deve-se avaliar a homogeneidade

dos dados para decidir sobre a manutenção da variável no decorrer do tratamento estatístico, evitando levar para o modelo variáveis com distribuição fortemente heterogênea.

Construção do modelo

Para a construção do modelo de regressão em si, os coeficientes "β" podem ser calculados por meio da função "lm" do pacote "stats" do R, a qual segue as formulações matemáticas do método dos mínimos quadrados ordinários.

Na sequência, as hipóteses da regressão devem ser testadas e validadas. No caso da hipótese i, utiliza-se o teste de hipóteses F, com nível de significância "$\alpha = 5\%$", que consiste em:

(a) H_0 (hipótese nula): $\beta_1, \beta_2, ..., \beta_j = 0$; e

(b) H_1 (hipótese alternativa): ao menos um parâmetro "β_j" é diferente de zero.

Os resultados do teste já fazem parte dos *outputs* da função "lm" do pacote "stats" do R. Dessa forma, rejeita-se "H_0" segundo o p-valor do seguinte modo:

(a) "p-valor $\leq 0,05$": indica que a correlação linear é estatisticamente significante; e

(b) "p-valor $> 0,05$": indica que a correlação linear não é estatisticamente significante.

De modo similar ao teste F de hipóteses, o teste t-Student deve ser aplicado para avaliar a significância de cada variável independente para o modelo de regressão. Se o p-valor do teste é menor que 0,05, a hipótese nula pode ser rejeitada e é possível afirmar que a variável é estatisticamente significante para o modelo, ao nível de confiança de 95%. Destaca-se que o modelo deve ser formado somente por variáveis independentes estatisticamente significantes. Os resultados do teste t-Student também integram os *outputs* da função "lm" do pacote "stats" do R.

Já a análise da hipótese II deve ser efetuada pelo fator de inflação de variância. Montgomery, Peck e Vining (2012) apontam que o fator em questão deve ser menor ou igual a 10 para que não haja multicolinearidade no modelo, ou seja, para que as variáveis independentes não possuam correlação linear entre si. O fator pode ser obtido por meio da função "vif" do pacote "car" do R.

Em relação à hipótese III, a normalidade pode ser analisada pelo teste de hipóteses de Shapiro-Wilk, valendo-se da função "shapiro.test" do pacote "stats" do R. O teste possui as seguintes hipóteses:

(a) H_0 (hipótese nula): os resíduos provêm de uma população com distribuição normal; e

(b) H_1 (hipótese alternativa): os resíduos não provêm de uma população com distribuição normal.

Caso a amostra tenha mais do que 30 observações, a normalidade deve ser investigada pelo teste de hipóteses de Shapiro-Francia, por meio da função "sf.test" do pacote "nortest" do R.

Por sua vez, a homocedasticidade pode ser investigada pelo teste de hipóteses de Breusch-Pagan, com auxílio da função "ncvTest" do pacote "car" do R. O teste de Breusch-Pagan é comumente utilizado para testar a hipótese nula de as variâncias dos resíduos serem constantes, contra a hipótese alternativa de elas serem função de algum regressor.

Tanto o teste de Shapiro-Wilk quanto o de Breusch-Pagan devem ser realizados com nível de significância "$\alpha = 5\%$".

Por fim, cumpre mencionar que, na eventualidade de o modelo não atender ao pressuposto de normalidade, é possível aplicar transformações à variável dependente para definir outras formas funcionais, que não a linear. Como exemplo, tem-se a transformação de Box-Cox (BOX; COX, 1964). Todavia, não faz parte do propósito deste artigo aprofundar nessa temática.

Avaliação da performance do modelo

O procedimento proposto para avaliar a performance é validação cruzada *k-fold*. Nesse tipo de validação, a amostra é dividida em "m" partes, sendo uma delas utilizada para testar o modelo e as demais para construí-lo. Tal procedimento é repetido até que cada parte da amostra seja utilizada para validação do modelo. Logo, "m" modelos são desenvolvidos e testados. Os erros obtidos para cada rodada são então utilizados para calcular a performance do modelo construído com todas as observações (*i.e.*, sem partição em "m" partes).

A presente metodologia adota um caso particular da validação *k-fold*, o método *leave-one-out*. Portanto, "m" sempre assume valor igual ao número de observações que compõem a amostra, sendo assim realizada uma estimativa de erro para cada observação retirada para validação.

O cálculo dos erros percentuais absolutos (EPAs) utilizados na validação *leave-one-out* segue a Equação 5.

$$EPA_i(\%) = \left|\frac{C_i - A_i}{A_i}\right| \times 100\% \quad \text{Eq. 5}$$

Onde:

C_i: custo estimado pelo modelo de regressão para a obra "i"; e

A_i: custo real da obra "i".

A medida de tendência central indicada para compilar os EPAs é a mediana, dado que ela é menos afetada por *outliers*, sobretudo em se tratando de amostras pequenas.

O processo *leave-one-out* pode ser desenvolvido com auxílio da função "trainControl" do pacote "caret" do R.

Modelagem: cadeiais públicas e penitenciárias

Com vistas a aplicar e testar a estrutura proposta nesta pesquisa, foi realizada uma modelagem com obras de construção de cadeias públicas e penitenciárias do Departamento Penitenciário Nacional do Ministério da Justiça e Segurança Pública (DEPEN/MJSP).

Dentre as definições iniciais de requisitos do modelo, cabe destacar as seguintes:

(a) contexto de aplicação: o modelo deve retratar o contexto brasileiro como um todo, em detrimento a unidades federativas (UFs) específicas;

(b) fase de projeto: estudo de viabilidade; e

(c) finalidade: criação de um referencial para a atividade de estimativa de custos de obras de construção de cadeias públicas e penitenciárias de segurança média.

Ao final da segunda fase proposta na estrutura, foram estabelecidas as sete variáveis independentes dispostas no Quadro 2.

Já a fase de coleta e elaboração do banco de dados foi realizada a partir dos arquivos de 27 obras de construção de cadeias públicas e penitenciárias disponibilizados pelo DEPEN/MJSP, que representam parte do acervo histórico de obras fomentadas pelo departamento. Essencialmente, foram fornecidos dois tipos de arquivo: projetos arquitetônicos e/ou planilhas orçamentárias.

Além dos dados das variáveis dependente e independentes, também foram coletadas informações sobre o percentual de benefícios e despesas indiretas (BDI) e o mês-base do orçamento.

Uma vez efetuada a coleta, os preços foram reajustados para um mesmo mês-base por meio do Índice Nacional de Custo da Construção (INCC), visando normalizar os efeitos decorrentes da inflação. Para tanto, utilizou-se a Equação 6:

$$P_{reaj} = P_{orça} \times \left[1 + \left(\frac{I_1 - I_0}{I_0}\right)\right] \quad \text{Eq. 6}$$

Onde:

P_{reaj}: preço reajustado, em reais;

$P_{orça}$: preço do orçamento, em reais;

I_1: índice publicado para o mês-base de reajuste (abril de 2020); e

I_0: índice publicado para o mês-base do orçamento.

Quadro 2 – **Variáveis independentes e respectivas métricas**

Variável	Métrica
Área construída	Metro quadrado
Quantidade de vagas	Unidade
Área total do terreno	Metro quadrado
Área interna à muralha	Metro quadrado
Perímetro da muralha	Metro
Quantidade de celas	Unidade
Desoneração da folha de pagamento	Sim ou não

Para o reajuste e normalização dos preços, adotou-se o mês-base de abril de 2020, que era a referência mais recente quando do desenvolvimento da modelagem.

Ainda em relação aos dados da variável dependente, cumpre mencionar que foram considerados apenas preços referentes ao orçamento da obra em si, sejam eles diretos ou indiretos. Então, descartaram-se preços relacionados à elaboração de projetos, realização de ensaios etc.

Cabe informar que, embora o DEPEN/MJSP tenha disponibilizado arquivos de 27 obras de construção de cadeias públicas e penitenciárias, o banco de dados foi composto por apenas 14 obras. Isso porque havia obras praticamente idênticas para uma mesma UF, de modo que sua incorporação na amostra poderia enviesar os resultados do tratamento estatístico, "atraindo" os coeficientes da regressão para sua direção. Como indicado nos requisitos do modelo, o estudo em questão tem o propósito de representar a realidade do Brasil como um todo, não apenas de um estado específico.

O conjunto de todas as informações mencionadas deu origem ao banco de dados utilizado na modelagem por regressão linear, o qual se encontra disponível para consulta na Tabela 3. Por motivos de confidencialidade, os dados referentes aos preços não puderam ser divulgados no corpo deste trabalho, assim como os nomes das cadeias e penitenciárias.

Nota-se que 43% da amostra proveio da região Norte, seguida das regiões Norte (21%), Sudeste (14%), Sul (14%) e Centro-Oeste (7%). Dessa forma, o espaço amostral foi considerado homogêneo para retratar o contexto geral brasileiro. Observa-se também que o mês-base mais antigo é junho de 2014, considerado relativamente recente para a aplicação em curso.

Na fase de modelagem dos dados, cabe pontuar alguns resultados do processo de análise exploratória. Em relação às variáveis "área interna à muralha" e "perímetro da muralha", verificou-se que nem todas as obras da amostra apresentaram a construção de muralha interna com passadiço, logo a variável pode não ser apropriada para representar todos os cenários, sendo assim inadequada para retratar as tipologias de obras em análise. Além do mais, a distribuição de frequências da variável "desoneração da folha de pagamento" revelou que 92% da amostra exibiu folha de pagamento desonerada, apontando elevada heterogeneidade da variável. Por fim, a investigação dos *boxplots* das variáveis quantitativas, associada à análise técnica de projetos e orçamentos das obras, culminaram na remoção da obra 14 da amostra. Ante o exposto, as variáveis "área interna à muralha", "perímetro da muralha" e "desoneração da folha de pagamento" foram desconsideradas, e a amostra foi reduzida para 13 obras.

Dado o tamanho amostral e respeitando os limites definidos por Park e Dudycha (1974), foram construídos modelos de regressão com no máximo duas variáveis independentes, assumindo "$R^2 = 0{,}50$", "$\varepsilon = 0{,}20$" e "$\gamma = 0{,}95$". Testaram-se todas as combinações possíveis entre as variáveis independentes restantes.

Tabela 3 – **Banco de dados**

UF	Obra	Área construída (m²)	Vagas (un)	Área total (m²)	Área interna à muralha (m²)	Perímetro da muralha (m)	Cela (un)	BDI (%)	Mês-base (SINAPI)	Desoneração da folha de pagamento (sim = 1; não = 0)
AP	1	8.254,62	422	37.899,67	18.268,13	548,91	64	25,22	08/19	1
AM	2	6.892,05	286	15.780,85	11.493,30	428,92		28,35	07/17	1
CE	3	6.763,62	168	20.357,12	14.380,75	480,10	104	24,98	12/17	1
GO	4	6.982,05	388	15.780,85	11.493,30	428,92	66	26,35	06/14	1
MA	5	6.219,11	314	20.577,09	0,00	0,00	56	25,00	09/17	1
MG	6	6.982,05	388	15.780,85	11.493,30	428,92	66		10/18	
PA	7	2.678,96	306	10.038,95	0,00	0,00	48	25,22	11/17	1
PB	8	13.964,10	748	31.561,70	22.986,60	857,84	134	28,63	01/18	1
RS	9	6.982,05	388	15.780,85	11.493,30	428,92	66	28,75	05/19	1
PI	10	6.359,94	336	26.096,65	0,00	0,00	105	26,43	04/19	1
RN	11	6.942,92	420		0,00	0,00	69	30,00	08/18	1
SC	12	10.862,91	364	22.511,24	17.224,11	529,40	79	25,00	09/18	0
SE	13	15.902,85	632	43.005,28	0,00	0,00	194	24,99	06/17	
SP	14	11.132,27	768					27,55	07/14	1

Ao final da modelagem, foram obtidos dois modelos recomendados para a estimativa de custos de obras de construção de cadeias públicas e penitenciárias de segurança média de interesse do DEPEN/MJSP. O primeiro deles foi construído com a variável "área construída", ao passo que segundo foi obtido com a variável "quantidade de vagas".

A avaliação da performance dos dois modelos foi realizada por meio da mediana dos erros percentuais absolutos obtidos pelo método *leave-one-out*, culminando em margens de erro de 25% para o primeiro modelo e 23% para o segundo.

As informações gerais de ambos os modelos são exibidas nas Tabelas 4 e 5, respectivamente. Já as Figuras 3 e 4 apresentam, respectivamente, os gráficos de dispersão construídos com as variáveis independentes dos modelos. Para manter o sigilo dos coeficientes, os modelos em si não puderam ser apresentados nesta pesquisa, tampouco puderam ser exibidos os valores referentes ao eixo da variável dependente.

Por meio das Tabelas 4 e 5, é possível verificar que ambos os modelos atenderam às hipóteses I e III da regressão linear, sendo que não é necessário avaliar a hipótese ii em modelos de regressão linear simples, pois só há uma variável independente. Tanto o modelo 1 como o 2 respeitaram os limites estipulados pela Prática Recomendada nº 17R-97 (ASSOCIATION..., 2011) para a estimativa de custos em nível de estudo de viabilidade, quando se comparam os valores das medianas dos erros percentuais absolutos (25 e 23%, respectivamente) com os limites para margem de erro apresentados na Tabela 1 (15 a 60%). Cabe destacar também o elevado coeficiente de determinação obtido para o modelo 2, indicando que 92% da variabilidade dos custos pode ser explicada pela quantidade de vagas.

Discussões

A estrutura proposta na seção do método de pesquisa pôde ser aplicada com sucesso ao cenário dos empreendimentos do Departamento Penitenciário Nacional do Ministério da Justiça e Segurança Pública.

Na fase de definição de requisitos do modelo, além do levantamento das informações mínimas apresentadas no princípio da modelagem, constatou-se a necessidade de alinhamento constante entre as partes interessadas para que o modelo atendesse ao propósito para o qual foi idealizado. Logo, a convergência de entendimentos deve ser buscada ao longo de toda a modelagem, visando minimizar o risco de o modelo final não ser utilizável para a estimativa de custos no contexto planejado.

Tabela 4 – **Modelo 1: variável "área construída"**

Critério	Dado
Mês-base	Abril de 2020
Tamanho da amostra	13 obras
R^2	0,55
Teste F (p-valor)	0,004
Teste de Shapiro-Wilk (p-valor)	0,193
Teste de Breusch-Pagan (p-valor)	0,203
Mediana dos EPAs (%)	25%
Limite inferior da área construída (m^2)	2.679
Limite superior da área construída (m^2)	15.903

Tabela 5 – **Modelo 2: variável "quantidade de vagas"**

Critério	Dado
Mês-base	Abril de 2020
Tamanho da amostra	13 obras
R^2	0,92
Teste F (p-valor)	0,000
Teste de Shapiro-Wilk (p-valor)	0,177
Teste de Breusch-Pagan (p-valor)	0,109
Mediana dos EPAs (%)	23%
Limite inferior da quantidade de vagas (un)	168
Limite superior da quantidade de vagas (un)	768

Figura 3 – **Gráfico de dispersão do modelo 1**

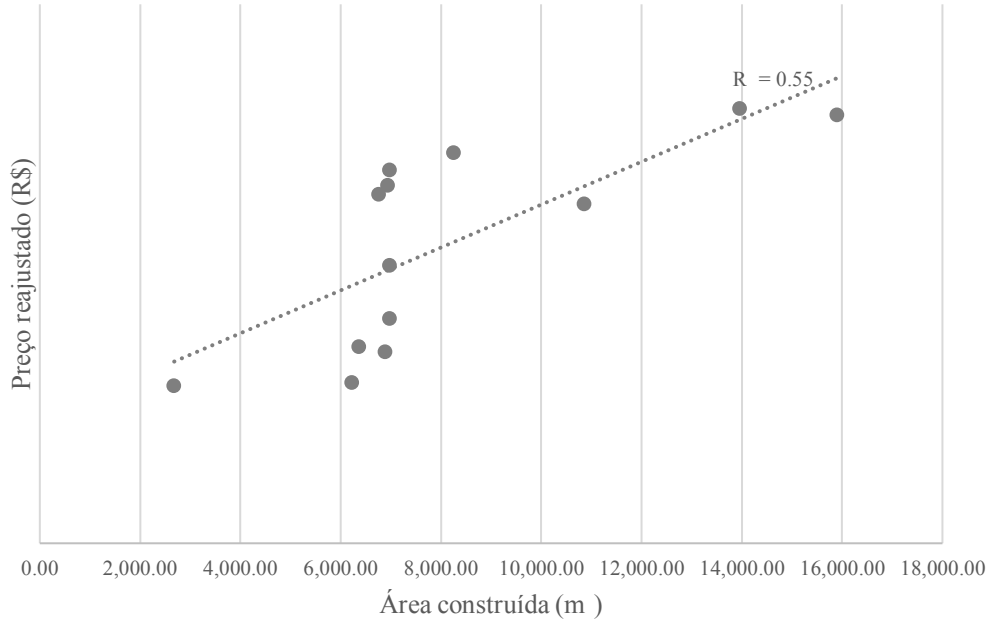
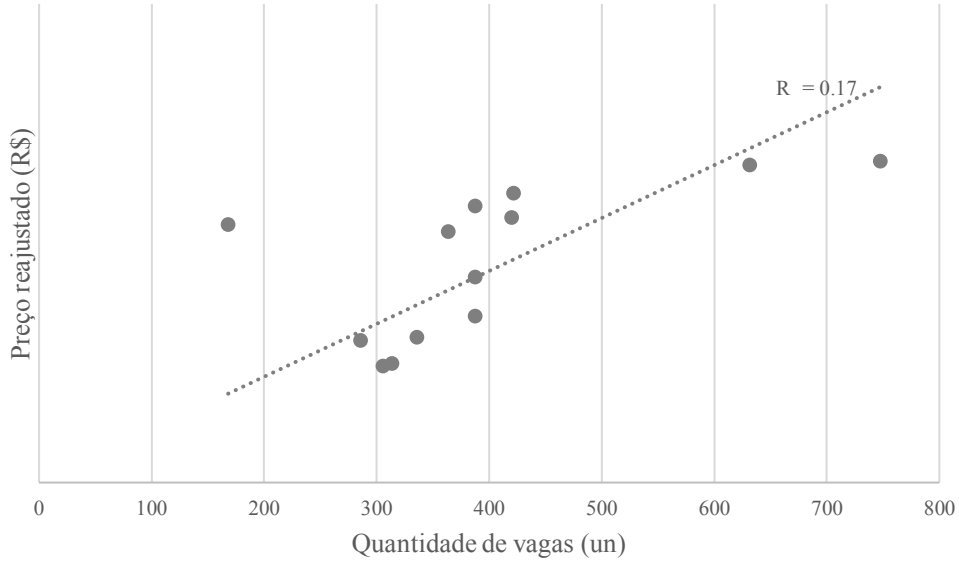


Figura 4 – **Gráfico de dispersão do modelo 2**



No decorrer da fase de seleção de variáveis independentes, surgiu a necessidade de elucidar a técnica da regressão linear, bem como os objetivos da modelagem, aos especialistas entrevistados para que eles pudessem julgar adequadamente as quatro proposições do questionário. O exemplo de questionário ilustrado na Figura 2 foi implantado em planilha editável, a qual foi distribuída para cada especialista. Isso trouxe agilidade ao cálculo do fator de influência da variável e mais eficácia à segunda fase da estrutura como um todo.

Em consonância com a literatura pesquisada (INTERNATIONAL..., 2008; GARDNER; GRANSBERG; JEONG, 2016), a fase de coleta e elaboração do banco de dados foi a que consumiu mais tempo (cerca de 2/3 do tempo) e recursos no ciclo da modelagem representado na Figura 1. Além do mais, no decorrer da fase em discussão, foram encontrados diversos entraves, dentro os quais se destacam:

(a) ausência de dados para as variáveis selecionadas;

(b) erros de digitação;

(c) falta de padrão entre as planilhas orçamentárias;

(d) informações divergentes entre projetos arquitetônicos e planilhas orçamentárias; e

(e) projetos em formato *pdf*, dificultando a coleta de variáveis quantitativas.

É importante salientar que, por mais complexas que as formulações matemáticas utilizadas para construir modelos baseados em dados sejam, elas não são capazes de balancear um banco de dados falho. Portanto, a fase em questão é um ponto crítico para o sucesso da modelagem proposta nesta pesquisa, pois, não apenas ela é a que consome mais tempo e recursos, mas também possui o potencial de inviabilizar a obtenção de modelos com performance satisfatória ou conferir viés aos resultados.

Quanto às fases de modelagem dos dados e de avaliação da performance do modelo, inicialmente cabe destacar que as funções do R utilizadas não demandaram alta capacidade de processamento e retornaram resultados com velocidade.

Em segundo lugar, ao final do processo de análise exploratória dos dados, a amostra foi reduzida para somente de 13 obras. Nesse ponto, cumpre mencionar que técnicas orientadas a dados requerem um espaço amostral representativo da população, seja em termos de qualidade da informação, seja em termos de quantidade de dados, para descrevê-la com boa precisão. Com base em dados do Conselho Nacional de Justiça, há 1.916 estabelecimentos penais no Brasil atualmente. Assumindo esse valor como o tamanho da população estudada, o erro máximo de estimação de 30% e o nível de significância de 5%, é possível verificar que o tamanho da amostra para estimar a média de uma população finita é inferior a 13. Logo, a amostra, apesar de pequena, pode ser considerada como representativa.

Ainda assim, cabem algumas considerações sobre o tamanho da amostra. No caso da regressão linear, modelos desenvolvidos a partir de amostras não representativas podem ser impactados de duas principais maneiras: apresentar coeficientes fortemente influenciados por uma única observação; exibir alta correlação linear entre as variáveis dependente e independentes, mas baixa capacidade preditiva.

O tamanho não representativo do espaço amostral também pode ter implicações de caráter prático. A extrapolação dos limites inferior e superior para os valores de entrada das variáveis independentes deve ser evitada, pois há incertezas quanto ao comportamento da variável dependente fora dos limites verificados para as variáveis independentes. Assim, amostras pequenas podem apresentar restrições de aplicação caso a amplitude dos limites não seja representativa da população.

Ainda em relação ao tamanho da amostra, puderam ser testados apenas modelos com duas variáveis independentes, em respeito às definições introduzidas por Park e Dudycha (1974). Após a fase de modelagem dos dados, os dois modelos de melhor performance foram obtidos com uma variável independente. Em ambos os casos, os limites da Prática Recomendada nº 17R-97 (ASSOCIATION..., 2011) foram atendidos e os pressupostos da regressão linear foram verificados. Logo, eles foram classificados como válidos para uso.

Quanto às margens de erro, observou-se que elas convergiram com maior parte dos estudos discutidos no referencial teórico que utilizaram a regressão linear como algoritmo. Por exemplo, Lowe, Emsley e Harding (2006), cujo objeto de estudo foi edificações, obtiveram um modelo de regressão com margem de erro de 25%. Mahamid (2011), que estudou obras de construção de rodovias, chegou a modelos com margens de erro entre 13 e 31%. Recentemente, Ogungbile, Rasak e Oke (2018), cujo tema também foi a construção de rodovias, e Elmousalami, Elyamany e Ibrahim (2018), que abordou a construção de canais de irrigação, desenvolveram modelos com erros de 8%. Cabe pontuar que os autores citados dispuseram de amostras significativamente maiores do que a utilizada na modelagem aplicada nesta pesquisa obras de construção de cadeias públicas e penitenciárias, e que seus modelos foram construídos com mais de uma variável independente, contribuindo para a melhor performance (em termos de precisão) verificada em alguns casos.

Considerações finais

As diretrizes elucidadas no método de pesquisa, em conjunto da estrutura apresentada na Figura 1, consistem na principal contribuição desta pesquisa para a construção de modelos para estimativa de custos de obras por meio da regressão linear, contemplando os elementos mínimos necessários à sua aplicação prática, tais como: fases e respectivos processos, formulações matemáticas, funções do software R Studio, indicadores, limites referenciais, dentre outras diretrizes. A estrutura proposta pode ser replicada pelos mais diversos atores, desde que atendida a premissa de haver uma série histórica de dados para fundamentar a modelagem, fomentando o uso de dados na indústria da construção civil como um todo.

Cabe destacar também, que diferentemente da ampla maioria das pesquisas que utilizam a técnica de regressão linear para fins de previsão, a estrutura aborda satisfatoriamente a importância de se verificar o atendimento aos pressupostos da regressão, sintetizando problemas decorrentes do não atendimento. Mais além, indica testes de hipóteses cabíveis à avaliação dos pressupostos.

Seguindo estritamente a estrutura delineada, foram desenvolvidos dois modelos válidos e com performance adequada à estimativa de custos das obras de construção de cadeias públicas e penitenciárias do DEPEN/MJSP. O modelo construído com a variável "área construída" obteve erro percentual absoluto mediano igual de 25%, ao passo que o erro do modelo gerado com a variável "quantidade de vagas" foi igual a 23%. Em ambos os casos, as margens de erro respeitaram os limites recomendados pela Prática Recomendada nº 17R-97 (ASSOCIATION..., 2011) e convergiram com algumas das pesquisas investigadas.

Uma vez que o uso dos modelos demanda somente informações sobre a área a ser construída ou a quantidade de vagas almejada, o Departamento Penitenciário Nacional do Ministério da Justiça e Segurança Pública, que dispõe dos coeficientes não divulgados nesta pesquisa por razões de confidencialidade, pode utilizar os modelos com praticidade e velocidade para fins de estimativa de custos em nível de estudo de viabilidade técnico-econômica. Desse modo, os modelos contribuem para o processo de tomada de decisão gerencial de empreendimentos penitenciários do departamento, bem como para sua previsão orçamentária e de investimentos.

A aplicação da estrutura proposta no contexto do DEPEN/MJSP atesta sua validade e retrata seu potencial de replicação por demais instituições públicas e privadas interessadas em utilizar a regressão linear para auxiliar processos internos de estimativa de custos de obras de construção civil.

Outro mérito deste estudo reside na proposição do fator de influência da variável. O indicador e os processos elucidados para seu cálculo sistematizam uma das fases mais subjetivas da modelagem por meio de regressão linear, que é a seleção de variáveis independentes. Naturalmente, a precisão do modelo na previsão dos custos está associada aos regressores escolhidos para compor a modelagem. Além do mais, a inclusão de inúmeras variáveis tende a aumentar os esforços para construção e uso do modelo, pois mais dados devem ser coletados e analisados.

No cenário brasileiro, o presente artigo estimula a promoção de discussões e estudos sobre o emprego de processos orientados a dados na estimativa de custos. Isso porque não foram identificadas pesquisas relacionadas ao tema em publicações nacionais ao longo da revisão da literatura realizada.

Em conclusão, acredita-se que modelos produzidos por meio da estrutura sejam capazes de trazer agilidade e simplicidade à estimativa de custos, bem como consistência e precisão. Acredita-se ainda que a introdução de técnicas consagradas e confiáveis, tais como a regressão linear, nos processos da construção civil tende a estimular a criação de uma cultura orientada por dados, um dos pressupostos da Indústria 4.0.

Referências

ASMAR, M. E.; HANNA, A. S.; WHITED, G. C. New approach to developing conceptual cost estimates for highway projects. **Journal of Construction Engineering and Management**, v. 137, n. 11, p. 942-949, 2011.

ASSOCIATION FOR THE ADVANCEMENT OF COST ENGINEERING. **Prática Recomendada Nº 17R-97**: Sistema de Classificação para Estimativa de Custos. Revisada em 29 de novembro de 2011.

BOX, G. E. P.; COX, D. R. An analysis of transformations. Journal **of the Royal Statistical Society, Series B**, v. 26, n. 2, p. 211-252, 1694.

ELMOUSALAMI, H. H. Artificial Intelligence and parametric construction cost estimate modeling: state-of-the-art review. **Journal of Construction Engineering and Management**, v. 146, n. 1, p. 03119008, 2020.

ELMOUSALAMI, H. H.; ELYAMANY, A. H.; IBRAHIM, A. H. Predicting conceptual cost for field canal improvement projects. **Journal of Construction Engineering and Management**, v. 144, n. 11, p. 04018102, 2018.

GARDNER, B. J.; GRANSBERG, D. D.; JEONG, H. D. Reducing data-collection efforts for conceptual cost estimating at a highway agency. **Journal of Construction Engineering and Management**, v. 142, n. 11, p. 04016057, 2016.

HASHEMI, S. T.; EBADATI, O. M.; KAUR, H. Cost estimation and prediction in construction projects: a systematic review on machine learning techniques. **SN Applied Sciences**, v. 2, n. 1703, 2020.

HEGAZY, T.; AYED, A. Neural network model for parametric cost estimation of highway projects. **Journal of Construction Engineering and Management**, v. 124, n. 3, p. 210-218, 1998.

HOLLAR, D. A. *et al*. Preliminary engineering cost estimation model for bridge projects. **Journal of Construction Engineering and Management**, v. 139, n. 9, p. 1259-1267, 2013.

INTERNATIONAL SOCIETY OF PARAMETRIC ANALYSTS. **Parametric estimating handbook**. 4th ed. Vienna, VA, 2008.

KIM, B-S. The approximate cost estimating model for railway bridge project in the planning phase using CBR method. **KSCE Journal of Civil Engineering**, v. 15, n. 7, p. 1149-1159, 2011.

KIM, B-S.; HONG, T. Revised case-based reasoning model development based on multiple regression analysis for railroad bridge construction. **Journal of Construction Engineering and Management**, v. 138, p. 154-162, 2012.

KIM, G-H.; AN, S-H.; KANG K-I. Comparison of construction cost estimating models based on regression analysis, neural networks, and case-based reasoning. **Building and Environment**, v. 39, n. 10, p. 1235-1242, 2004.

LIU, M. *et al*. Preliminary engineering cost-estimation strategy assessment for roadway projects. **Journal of Management in Engineering**, v. 29, n. 2, p. 150-157, 2013.

LOWE, D. J.; EMSLEY, M. W.; HARDING, A. Predicting construction cost using multiple regression techniques. **Journal of Construction Engineering and Management**, v. 132, n. 7, p. 7850-758, 2006.

MAHALAKSHMI, G.; RAJASEKARAN, C. Early cost estimation of highway projects in india using artificial neural network. **Sustainable Construction and Building Materials**, v. 25, p. 659-672, 2019.

MAHAMID, I. Early cost estimating for road construction projects using multiple regression techniques. **Australasian Journal of Construction Economics and Building**, v. 11, n. 4, p. 87-101, 2011.

MONTGOMERY, D. C.; PECK, E. A.; VINING, G. G. **Introduction to linear regression analysis**. 5th. ed. Hoboken: John Wiley & Sons, 2012.

MOSELHI, O.; SIQUEIRA, I. Neural networks for cost estimating of structural steel buildings. **AACE International Transactions**, p. IT22, 1998.

OGUNGBILE, A.; RASAK, K.; OKE, A. E. Developing cost model for preliminary estimate of road projects in Nigeria. **International Journal of Sustainable Real Estate and Construction Economics**, v. 1, n. 2, p. 182-199, 2018.

PARK, C. N.; DUDYCHA, A. L. A cross-validation approach to sample size determination for regression models. **Journal of the American Statistical Association**, v. 69, n. 345, p. 214-218, 1974.

PETROUTSATOU, C.; LAMBROPOULOS, S.; PANTOUVAKIS, J-P. Road tunnel early cost estimates using multiple regression analysis. **Operational Research**, v. 6, n. 3, p. 311-322, 2006.

PETRUSEVA, S. *et al*. Construction costs forecasting: comparison of the accuracy of linear regression and support vector machine models. **Tehnicki Vjesnik-Technical Gaz**, v. 24, n. 5, p. 1431-1438, 2017.

RAFIEI, M. H.; ADELI, H. Novel machine-learning model for estimating construction costs considering economic variables and indexes. **Journal of Civil Engineering and Management**, v. 144, n. 12, p. 04018106, 2018.

ROSTAMI, J. *et al*. Planning level tunnel cost estimation based on statistical analysis of historical data. **Tunnelling and Underground Space Technology**, v. 33, p. 22-33, 2013.

SANDERS, S. R.; MAXWELL, R. R.; GLAGOLA, C. R. Preliminary estimating models for infrastructure projects. **Cost Engineering**, v. 34, n. 8, p. 7-13, 1992.

SONMEZ, R. Parametric range estimate of building costs using regression models and bootstrap. **Journal of Construction Engineering and Management**, v. 134, n. 12, p. 1011-1016, 2008.

SONMEZ, R.; ONTEPELI, B. Predesign cost estimation of urban railway projects with parametric modeling. **Journal of Civil Engineering and Management**, v. 15, n. 4, p. 405-409, 2009.

WILLIAMS, T. P. Predicting final cost for competitively bid construction projects using regression models. **International Journal of Project Management**, v. 21, n. 8, p. 593-599, 2003.

ZHAI, K.; JIANG, N.; PEDRYCZ, W. Cost prediction method based on an improved fuzzy model. **International Journal of Advanced Manufacturing Technology**, v. 65, p. 1045-1053, 2012.

ZHANG, Y.; MINCHIN JUNIOR, R. E.; AGDAS, D. Forecasting completed cost of highway construction projects using LASSO regularized regression. **Journal of Construction Engineering and Management**, v. 143, n. 10, p. 04017071, 2017.

ZHU, W-J.; FENG, W-F.; ZHOU, Y-G. The application of genetic fuzzy neural network in project cost estimate. **International Conference on E-Product E-Service and E-Entertainment**, New York, 2010.

Leandro Modesto Prates Beltrão
Programa de Pós-Gradução em Estruturas e Construção Civil, Departamento de Engenharia Civil e Ambiental | Universidade de Brasília | Campus Universitário Darcy Ribeiro | Brasília – DF – Brasil | CEP 70910-900 | Tel.: (61) 99691-5169 | E-mail: leandromodesto.eng@gmail.com

Michele Tereza Marques Carvalho
Programa de Pós-Gradução em Estruturas e Construção Civil, Departamento de Engenharia Civil e Ambiental | Universidade de Brasília | Tel.: (61) 3107-1010 | E-mail: micheletereza@gmail.com

Raquel Naves Blumenschein
Parque de Inovação e Sustentabilidade do Ambiente Construído, Faculdade de Arquitetura e Urbanismo | Universidade de Brasília | Campus Universitário Darcy Ribeiro | Brasília – DF – Brasil | CEP 70910-900 | Tel.: (61) 3107-7429 | E-mail: raquelblum@terra.com.br

Álvaro Teixeira de Paiva
Instituto Brasileiro de Economia | Fundação Getulio Vargas | SGAN, Quadra 602, Asa Norte | Brasília – DF – Brasil | CEP 70830-051 | Tel.: (61) 98198-7622 | E-mail: alvaro.paiva95@gmail.com

Maíra Vitoriano Rodrigues de Freitas
Empresa de Planejamento e Logística | Edifício Parque Cidade Corporate, Torre C, Via W4 Sul, Lote C, Asa Sul | Brasília – DF – Brasil | CEP 70308-200 | Tel.: (61) 98286-0926 | E-mail: mairavitoriano@gmail.com

Ambiente Construído
Revista da Associação Nacional de Tecnologia do Ambiente Construído
Av. Osvaldo Aranha, 99 - 3º andar, Centro
Porto Alegre – RS - Brasil
CEP 90035-190
Telefone: +55 (51) 3308-4084
www.seer.ufrgs.br/ambienteconstruido
www.scielo.br/ac
E-mail: ambienteconstruido@ufrgs.br