EPiC Series in Built Environment
Volume 3, 2022, Pages 398–406
ASC2022. 58th Annual Associated Schools
of Construction International Conference

Feature Engineering for development of a Machine Learning Model for Clash Resolution

Ashit Harode and Walid Thabet, Ph.D., CM-BIM
Virginia Tech
Blacksburg, Virginia

Fernanda L. Leite, Ph.D., P.E., M.ASCE
University of Texas at Austin
Austin, Texas

To automate clash resolution tasks, it is important to capture domain knowledge for the Machine Learning (ML). One way to add domain knowledge is by training data that divides tasks into input and output variables. The selection of input variables that are most relevant to a task is an important step towards automation. In this paper, the authors detail framework that uses literature review, industry interviews, and Modified Delphi to capture domain knowledge for clash resolution. The features identified through this paper can in future be processed through Feature Selection, that can provide empirical evidence of why the selected features or set of features are important to ML algorithm. Data collection processes discussed in this paper is not finalized and is discussed to help provide readers with framework of the proposed systematic method. Factors considered when resolving clashes were identified through literature review (22 factors) and industry interviews (16 factors). 14 factors identified from the interviews had a similar matching factor in the literature reviewed, the other 2 factors were not mentioned in any publications found during the initial literature review. After comparing results from literature review and interviews, 13 factors were considered critical for automating clash resolution.

Key Words: Feature Engineering, Clash Coordination, Machine Learning, Clash Resolution, Dimensionality Reduction, Features and Labels.

Introduction and Objective

Design coordination, including clash identification and resolution, is a construction task that can benefit from automation through ML. Hsu, Chang, Chen, and Wu (2020) have explored the use of supervised learning to automate the resolution of clashes occurring between mechanical components with considerable accuracy. However, due to the limited availability of a training data set, the model developed could be subject to over-fitting reducing the accuracy of the model for new clashes. This limitation can be overcome by providing a larger training dataset. However, labeling the training dataset highly relies on human experts and labor work (Huang & Lin, 2019).

T. Leathem, W. Collins and A. Perrenoud (eds.), ASC2022 (EPiC Series in Built Environment, vol. 3), pp. 398–406

To overcome this limitation, Harode and Thabet (2021) proposed a combined supervised-reinforcement ML approach to develop an automation model with high accuracy using a limited training dataset. The proposed model uses supervised learning with a limited labeled training dataset as pre-training for a reinforcement learning component. Reinforcement learning would use the supervised learning model to interact with clashes in a BIM model and improve on the automation model with each iterative interaction. The supervised learning component of the proposed model requires domain knowledge in the form of training dataset as input. Supervised learning algorithms develop new knowledge and make decisions by relying on accurate and complete labeled training datasets related to the business problem at hand. A labeled dataset contains both the input variables and corresponding output variables. In supervised learning, the input variables are referred to as features and the output variables are referred to as labels. Selecting features that appropriately depict the ML task is an important step towards adding domain knowledge to ML algorithms (Sutton & Barto, 2018). To generate optimal results from the model, it is important that training data used as input must contain features that are most relevant to the task (Theobald, 2017).

The objective of this paper is to present a framework that can be used to identify and capture decision-making data used by industry experts during their clash resolution process to extract the required features for the proposed combined supervised reinforcement learning model. Literature review and partial industry interviews are conducted in this paper to identify and capture data on factors used by industry experts' to resolve clashes as part of domain knowledge collection step of Feature Engineering. In future, Feature Engineering techniques like Feature Selection, can be applied to selected features to provide evidence supporting better efficacy of ML algorithm. Along with manipulation and transformation of selected features to facilitate development of ML model using selected features. As this paper discusses topics and keywords related to the field of ML, to facilitate the reading if the article by a larger audience Table 1 has defined these keywords.

Table 1

Keywords related to ML

Keywords	**Definition**
ML	A subfield of computer science that provides computers with the ability to learn without explicit programming (Theobald, 2017)
Supervised Learning	A ML technique that analyzes combination of known inputs and outputs to predict output to future inputs (Theobald, 2017).
Reinforcement Learning	A ML technique that develops automation knowledge by randomly interacting with the tasks to achieve desired output (Theobald, 2017).
Training Dataset	Known combination of input and output variable that can be used to develop automation model through supervised learning
Labeled Dataset	A dataset that contains both the inputs and their corresponding outputs.
Automation Model	An algorithmic equation developed through the ML process that facilitates in the automation of task.
Features	Inputs that are used to describe the data points.
Labels	Desired outputs corresponding to the given features (inputs).
Feature Engineering	"Act of extracting features from raw data and transforming them into format suitable for ML" (Zheng & Casari, 2018).
Feature Selection	Statistics-intensive process that provides empirical evidence on why certain features or set of features are important for automation (Ramasubramanian & Singh, 2019).

Domain Knowledge	Knowledge specific to the specialized discipline.
Over-fitting	Statistical error in ML where the developed model closely aligns with the limited data points.

The following sections will discuss the importance of Feature Engineering for automation of clash resolution, the methodology adopted to identify factors considered by industry experts for clash resolution, and the final set of 13 factors selected to automate clash resolution using the proposed ML model. The paper concludes with a discussion on the results of data collection and analysis and proposed future research work.

Literature Review

In ML, a feature is used to describe a data point (Dong & Liu, 2018). When performing clash resolution, for example, the system type of the clashing elements can be considered features of the clash. For effective ML, features that help in accurately describing/defining the task that needs to be automated should be identified and selected. For a given task, the process of selecting, formulating, and transforming the most appropriate features is called Feature Engineering (Zheng & Casari, 2018). A robust Feature Engineering process can provide ML algorithms with the following benefits: (1) improved predictive performance of the ML model, (2) faster and computationally less heavy ML process, (3) develop a better understanding of data relationships, and (4) create an explainable and implementable ML model (Ramasubramanian & Singh, 2019).

The concept of Feature Engineering can be divided into two supporting processes: (1) Business/domain knowledge, and (2) Feature Selection (Ramasubramanian & Singh, 2019). The domain knowledge process focuses on making sure that the features selected make sense and accurately reflect the domain knowledge. While Feature Selection is a more statistical-intensive process focused on providing empirical evidence to support the selection of a feature or set of features for a ML algorithm. In this paper, the authors focus on using Feature Engineering to collect and analyze domain knowledge for the automation of clash resolution.

Korman, Fischer, and Tatum (2003) captured knowledge related to the design, construction, operations, and maintenance of MEP systems using a research project to create a computer tool that can assist in resolving MEP coordination problems. Radke, Wallmark, and Tseng (2009) investigated an alternate approach for clash detection and resolution based on the parametric description of design elements. Wang and Leite (2016) proposed a systematic way to capture clash features and associated solutions for MEP coordination to support clash documentation. Results from their research assisted in the management of clash coordination and allowed the capture of existing domain knowledge to support future decision-making. Several research efforts are being expended to use ML for the automation of clash coordination. Hsu et al. (2020) used six features to define a clash and develop a supervised learning model to automate clash resolution of a student residence basement. Huang and Lin (2019) also performed a Feature Selection study to select six features to automate the classification of clashes using supervised learning. Hu, Castro-Lacouture, and Eastman (2019) analyzed six kinds of spatial relationships between clashing elements to develop a spatial network to eliminate irrelevant clashes, select clashes without enough room, and select clashes where the movement of one object can resolve multiple clashes. Hu, Castro-Lacouture, Eastman, and Navathe (2020) also designed an optimization algorithm to determine the optimal sequence for clash correction. This algorithm was based on clashing volume, the impact of moving one clashing element on the other MEP element in its proximity, the relationship between the clashing elements, and logical connection relations between building components.

Based on the review of the literature, the authors have identified two major research gaps. The research focused on formalizing knowledge for clash coordination focused only on documenting clash cases and knowledge associated with clash resolution but did not address the use and formatting of the knowledge collected as input to ML models. Literature focused on automating clash coordination did not address the methodology of identifying the features used in their automation models or why these features were selected. To overcome these knowledge gaps, the authors in this paper propose a systematic methodology using Feature Engineering to extract domain knowledge for automation of clash resolution. Using this methodology, the authors define a list of factors that should be considered as features for the proposed combined ML algorithm.

Research Approach

The process of Feature Engineering includes selecting, transforming, and formulating features that are most appropriate for the automation of a given task. In this paper the authors focus on the selection of features that most accurately reflects industry professionals' decision making for clash resolution, i.e., domain knowledge for clash resolution. Figure 1 details the entire proposed Feature Engineering step for the development of a strong ML model for the automation of clash resolution. In this paper, the authors have focused only on the domain knowledge collection process of Feature Engineering.

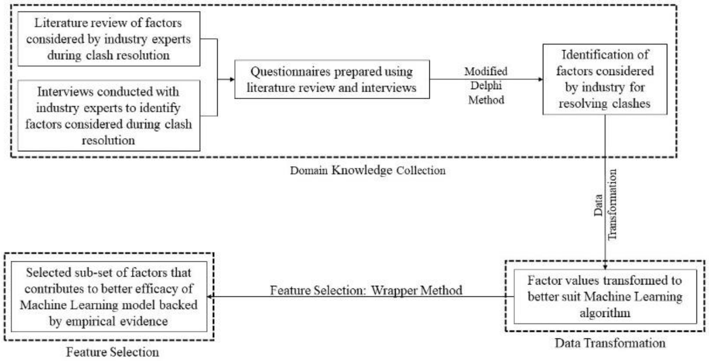
Literature review of factors considered by industry experts during clash resolution
Interviews conducted with industry experts to identify factors considered during clash resolution
Questionnaires prepared using literature review and interviews
Modified Delphi Method
Identification of factors considered by industry for resolving clashes
Domain Knowledge Collection
Data Transformation
Factor values transformed to better suit Machine Learning algorithm
Data Transformation
Feature Selection: Wrapper Method
Selected sub-set of factors that contributes to better efficacy of Machine Learning model backed by empirical evidence
Feature Selection

Figure 1. Complete process of Feature Engineering proposed to develop a ML model.

A Feature Engineering process to collect domain knowledge will comprise three main steps:

1. **Literature Review:** The first step involved the review of literature focused on clash resolution automation and industry best practices for clash coordination, summarized in Table 2. Based on the literature reviewed, 22 factors that should be considered in resolving clashes were identified. These factors were grouped into 4 categories: (1) Geometric Characteristic: Properties of the clashing element describing their geometry, (2) Functional Characteristic: Properties of the clashing elements related to their function, installation, operation, and maintenance,(3) Topological Relation: How the elements clash and how they intersect (Hu et al., 2019), and (4) Spatial Relation: Spatial relationship of the clashing elements relative to the surrounding elements.

Table 2
Factors identified through literature review

Factor Category	Factor	Factor Description	Literature
1. Geometric/Physical Characteristic	Start and End Point (X, Y, Z)	The 3D coordinates of the endpoints of the clashing elements	(Korman et al., 2003; Radke et al., 2009)
	Clash Component Dimensions	The height, width, length, and radius of the clashing elements.	(Korman et al., 2003; Radke et al., 2009; Wang & Leite, 2016)
	Baseline direction of the clashing elements	The geometric baseline direction of clashing elements (e.g., Horizontal and Vertical)	(Hsu et al., 2020)
	Element Slope	The existing slope of the clashing elements.	(Wang & Leite, 2016)
2. Functional/Operational Characteristic	Clashing element's type	The type of each clashing element (e.g., pipes, duct, structural framing)	(Huang & Lin, 2019; Wang & Leite, 2016)
	Clashing element's system type	The building system each clashing element belongs to	(Hsu et al., 2020; Hu et al., 2020; Korman et al., 2003; Wang & Leite, 2016)
	Constrained Slope	The required slope that needs to be maintained for the clashing elements	(Wang & Leite, 2016)
	Insulation	The size of insulation present around the clashing elements	(Wang & Leite, 2016)
	Clashing Element's Material	The material of clashing elements	(Wang & Leite, 2016)
	Rigidity of the clashing elements	Clashing element rigid or flexible?	(Wang & Leite, 2016)
	Critical Element in the clash	The presence of a critical element in the clash	(Wang & Leite, 2016)
3. Topological Relation	Hard or Soft Clash	Is the clash hard or soft?	(Hu et al., 2019; Radke et al., 2009; Wang & Leite, 2016)
	Clash Distance	The relative distance of the clashing elements	(Huang & Lin, 2019)
	Clash Type	Clashing elements oriented parallel or perpendicular to each other	(Hsu et al., 2020; Hu et al., 2019)
	Intersection Type	Intersection of the clashing elements is penetrating or puncturing	(Hsu et al., 2020; Hu et al., 2019)
	Clashing Volume	The overlapping volume of the clashing elements	(Hu et al., 2020; Wang & Leite, 2016)
	Clash Group	Group of clashes involving a common clashing element	(Wang & Leite, 2016)
4. Spatial Relation	Moveable Area	Area where the element with low priority can move without violating any space constraints	(Hu et al., 2019; Radke et al., 2009; Wang & Leite, 2016)
	Location of the clash	The floor/room where the clash is located (e.g., Mechanical Room)	(Huang & Lin, 2019; Korman et al., 2003; Wang & Leite, 2016)
	Clash Point	3D coordinates of the clash location	(Huang & Lin, 2019)
	Impacted Object	Object not part of the clash but is present within proximity of the clash	(Hu et al., 2019; Korman et al., 2003)
	Connections	Number of vertical and horizontal connections/fittings per length of the clashing element	(Korman et al., 2003)

2. **Industry Interviews:** Unstructured/structured interviews with industry experts from various disciplines (GC and mechanical) were conducted to augment the findings of the literature review and fill any gaps. Additionally, these interviews helped the authors to reduce redundancy and eliminate any overlaps between the factors that were identified during the literature review (Gunduz & Elsherbeny, 2020). The interviews focused on discovering answers to several questions, including: (1) What factors does the project coordination team consider while resolving a clash?, (2) What considerations do industry professionals make when resolving a clash?, (3) How does the orientation of clashing elements affect the clash resolution?, and (4) How are priorities between clashing elements established? Interviewees were also asked to discuss specific examples of clashes from BIM models provided and analyzed by the authors prior to the interviews The interviews were video recorded to facilitate further analysis of the discussion following the meetings. All interviews were transcribed and coded by the authors to identify specific clash resolution factors discussed during each interview.

3. **Industry Survey using Modified Delphi:** Through the interviews and literature review process, a substantial list of factors considered by design coordination teams while resolving a clash will be generated. This third step will focus on developing a common consensus on the identified factors by the industry experts. To achieve this objective, a Modified Delphi Method will be used. The intended purpose of the Delphi methodology is to obtain a common consensus of qualified industry experts on a particular subject (factors considered for clash resolution) by allowing them to look at a set of updated questionnaires along with feedback provided. If the questionnaire is developed through literature review and interviews, this process is called modified Delphi (Gunduz & Elsherbeny, 2020). A survey will be sent to industry experts listing the identified clash resolution factors. Each individual receiving the survey results will be asked to select all the factors that they believe affect clash resolution decisions. Once the response for the first round of the questionnaires has been received, the result of the first round will be analyzed. This analysis will be sent back to industry experts along with the first-round of questionnaire. In the second round, industry experts will be given the opportunity to change their answers based on the analysis summary. They can choose to keep their original answer or modify it. Individuals who decide to keep their original answer will be asked to provide an explanation. Industry experts will be allowed to add any factors they think are important for clash resolution but are missing from the given list. Multiple rounds of modified Delphi will be conducted until a common consensus is reached.

Results

Table 3a shows a comparison between factors identified from literature review and interviews, factors that were not discussed are omitted from the table. Using this comparison, a final preliminary list of 13factors considered by the authors for automation of clash resolution was prepared and shown in Table 3b.

Table 3a
Comparison between literature reviews and interviews

Factors identified through Literature Review	Factors discussed in Interview 1	Factors discussed in Interview 2	Factors discussed in Interview 3
Clash Component Dimensions.	X	X	X
Baseline Direction of the Clashing Elements.		X	
Element Slope.	X	X	X

Clashing Element’s Type.	X	X	X
Clashing Element’s System Type.	X	X	X
Constrained Slope.		X	X
Insulation.	X		
Clashing Element’s Material.			X
Rigidity of the Clashing Elements.			X
Critical Element in the Clash.		X	X
Clash Group.		X	
Moveable Area.	X	X	X
Location of the Clash.	X	X	X
Connections.	X	X	X
Additional Factors discussed during Interview 1 and 3: Cost of resolving a clash			
Additional Factors discussed during Interview 3: Construction stage the clashing elements is in			

Table 3b

List of factors considered while resolving the clashes

S. No.	Factors	S. No.	Factors
1.	Start and End Point (X, Y, Z)	8.	Rigidity of the Clashing Elements
2.	Clash Component Dimensions	9.	Critical Element in the Clash
3.	Clashing Element’s Type	10.	Clash Group
4.	Clashing Element’s System Type	11.	Moveable Area
5.	Constrained Slope	12.	Location of the Clash
6.	Insulation	13.	Connections
7.	Clashing Element’s Material		

Three rounds of initial interviews were conducted with industry experts experienced in clash coordination. These interviews were coded, and factors considered while resolving clashes were identified by the authors. Out of the 22 factors identified in the literature reviewed, only 14 matching factors were identified from the interviews. Factors related to topological relation between the clashing element were not discussed during the interview, except for the ‘Clash Group’ factor. Other factors that were not discussed during the interview were ‘Start and End Point (X, Y, Z)’, ‘Clash Point’, and ‘Impacted Object’. The ‘Start and End Point (X, Y, Z)’ factor, although not discussed during the interviews, was still included in the final list of factors in place of the ‘Baseline Direction’ factor. The authors argue that the baseline direction of clashing elements is a function of the coordinates of the endpoints of the elements which can be easily obtained using Navisworks or Revit Application Programming Interface. Another reason for selecting the ‘Start and End Point (X, Y, Z)’ factor over element baseline direction is endpoint coordinates can also help represent the current location and orientation of the clashing element in the ML algorithm, hence providing more useful and necessary information. Using the ‘Start and End Point (X, Y, Z)’ as a factor can also eliminate the use of the ‘Element Slope’ factor in the ML algorithm. As the existing slope of the element is also going to be the function of its endpoint coordinates. Using ‘Start and End Point (X, Y, Z)’ allows for removing the ‘Element Slope’ as a feature, making the ML algorithm computationally less heavy without removing the influence of the ‘Element Slope’ factor from the overall decision making.

Discussion

All clash resolution strategies discussed during the industry interviews focused on looking at the clash and its surrounding, and other connected model elements. Topological relations describing a single clash were not discussed so far. Factors such as 'Clash Point' and 'Impacted Object' were not discussed in the interviews conducted but can be considered as a part of the factor 'Moveable Area' which relates to the area and elements surrounding the clash. Another factor that was not discussed during the interviews was the 'Start and End Point (X, Y, Z)', Industry experts did not see the need to understand the orientation and baseline direction of the clashing element to keep the updated elements as close to the original design as possible while resolving the clash. The authors concluded that orientation and baseline direction of the clashing elements can become the function of their start and endpoint coordinates in ML algorithm, 'Start and End Point (X, Y, Z)' can be used to replace baseline direction and orientation of elements as factors for the ML algorithm. Another aspect of Feature Engineering that was discussed in this paper was reducing the dimensionality of the ML problem to make it computationally less heavy without eliminating the influence of necessary factors. As the existing slope of the clashing element can also be expressed using the endpoint coordinates of the clashing element. Using 'Start and End Point (X, Y, Z)' as one of the factors in our ML algorithm will eliminate the necessity of using 2 additional factors ('Element Slope' and 'Baseline Direction') without compromising their influence on the algorithm. During the interviews, two other factors that influence the clash resolution but were not found in the literature were also identified. Industry experts during the interviews discussed the consideration of the cost of resolving clashes. For example, in a clash scenario discussed related to a conflict between a duct and a pipe, an argument for modifying the duct location (usually given a higher priority) was made instead of the pipe (usually with lower priority). The reason for this exception was attributed to the pipe's copper material that requires more labor cost to modify and more costly to add additional copper fittings to resolve the clash. Making changes to the copper pipe, in this case, would have increased project cost. The authors suggest that the cost of clashing elements while resolving the clash should be considered as one of the factors and will be added as a feature in the proposed ML automation model. Another factor discussed during the interviews was the stage at which the clashing BIM element is, in the construction process. If a pipe has already been prefabricated for installation and a clash is identified that includes the pipe, it would be more cost and time-effective to move the other clashing elements. Through these interviews, it was realized that initial hierarchy of priority defined for different elements should sometimes be modified, when situations involving elements of lower priority may have a higher cost impact over elements of higher priority.

Concluding Remarks

The authors through the literature review and interviews have identified 13 critical factors that the industry experts consider while resolving clashes. These factors represent the domain knowledge utilized by industry professionals to make decision on clash resolution. As the next step within Feature Engineering, in future the authors would like to focus on data transformation to match a format suitable for ML algorithm. For example, text-based value of clashing element's system type can be converted to their numeric OmniClass values, "HVAC Ducts and Casings" to "22233100". Another step within Feature Engineering is Feature Selection, the authors plan to use wrapper methods for Feature Selection to evaluate the performance of the predictive algorithm for different subset of the factors and select the subset of factors that generates the most accurate predictive performance (Chandrashekar & Sahin, 2014). Once the Feature Engineering step is completed the selected factors will be utilized in future research to input domain knowledge to a proposed combined supervised-reinforcement ML algorithm that can more efficiently automate clash resolution based on industry standards. The goal of this research was to provide a framework to identify factors considered by industry experts while resolving clashes as part of domain knowledge collection process of Feature Engineering for the automation of clash resolution. The Feature Engineering process proposed in this paper focused on selections of factors that served two purposes, (1) accurate transfer of domain knowledge to the ML algorithm, (2) assist in making ML algorithm computationally less heavy without compromising its efficiency. In this paper, the authors have only completed the literature review step of the proposed Feature Engineering methodology. The initial interviews conducted in this research utilized two construction industry experts. As part of future work, the authors plan to conduct more interviews using a bigger pool of experts from different companies. Once the interviews are completed, the authors plan to update the list of factors. Following additional interviews, a Modified Delphi Process is planned to determine consensus among industry experts, hence concluding the domain knowledge collection process of Feature Engineering.

Reference

Chandrashekar, G., & Sahin, F. (2014). A survey on feature selection methods. Computers & Electrical Engineering, 40(1), 16-28. doi:https://doi.org/10.1016/j.compeleceng.2013.11.024

Dong, G., & Liu, H. (2018). Feature engineering for machine learning and data analytics: CRC Press.

Gunduz, M., & Elsherbeny, H. A. (2020). Operational framework for managing construction-contract administration practitioners' perspective through modified Delphi method. JOURNAL OF CONSTRUCTION ENGINEERING AND MANAGEMENT, 146(3), 04019110.

Harode, A., & Thabet, W. (2021). Investigation of Machine Learning for Clash Resolution Automation. EPiC Series in Built Environment, 2, 228-236.

Hsu, H.-C., Chang, S., Chen, C.-C., & Wu, I. C. (2020). Knowledge-based system for resolving design clashes in building information models. Automation in Construction, 110, 103001. doi:https://doi.org/10.1016/j.autcon.2019.103001

Hu, Y., Castro-Lacouture, D., & Eastman, C. M. (2019). Holistic Clash Resolution Improvement Using Spatial Networks. In Computing in Civil Engineering 2019: Visualization, Information Modeling, and Simulation (pp. 473-481): American Society of Civil Engineers Reston, VA.

Hu, Y., Castro-Lacouture, D., Eastman, C. M., & Navathe, S. B. (2020). Automatic clash correction sequence optimization using a clash dependency network. Automation in Construction, 115, 103205.

Huang, Y., & Lin, W. Y. (2019, 2019). Automatic Classification of Design Conflicts Using Rulebased Reasoning and Machine Learning-An Example of Structural Clashes Against the MEP Model.

Korman, T. M., Fischer, M. A., & Tatum, C. B. (2003). Knowledge and Reasoning for MEP Coordination. JOURNAL OF CONSTRUCTION ENGINEERING AND MANAGEMENT, 129(Part 6), 627-634.

Radke, A., Wallmark, T., & Tseng, M. (2009). An automated approach for identification and resolution of spatial clashes in building design. Paper presented at the 2009 IEEE International Conference on Industrial Engineering and Engineering Management.

Ramasubramanian, K., & Singh, A. (2019). Machine learning using R : with time series and industry-based uses in R [1 online resource (712 pages)](Second edition. ed.). Retrieved from https://doi.org/10.1007/978-1-4842-4215-5

Sutton, R. S., & Barto, A. G. (2018). Reinforcement learning: An introduction: MIT press.

Theobald, O. (2017). Machine learning for absolute beginners: a plain English introduction: Scatterplot press.

Wang, L., & Leite, F. (2016). Formalized knowledge representation for spatial conflict coordination of mechanical, electrical and plumbing (MEP) systems in new building projects. Automation in Construction, 64, 20-26.

Zheng, A., & Casari, A. (2018). Feature engineering for machine learning: principles and techniques for data scientists: " O'Reilly Media, Inc.".