CIRP Annals - Manufacturing Technology 72 (2023) 385–388

Contents lists available at ScienceDirect

CIRP Annals - Manufacturing Technology

journal homepage: https://www.editorialmanager.com/CIRP/default.aspx

Cognitive data imputation: Case study in maintenance cost estimation

John Ahmet Erkoyuncu (2)[a,*], Bernadin Namoano[a], Dominik Kozjek[b], Rok Vrabič (2)[b]

[a] Cranfield University, School of Aerospace, Transport and Manufacturing, United Kingdom
[b] University of Ljubljana, Faculty of Mechanical Engineering, Slovenia

ARTICLE INFO

Article history:
Available online 28 April 2023

Keywords:
Artificial intelligence
Maintenance
Cost estimation

ABSTRACT

Cost estimation is critical for effective decision making in engineering projects. However, it is often hampered by a lack of sufficient data. For this, data imputation techniques can be used to estimate missing costs based on statistical estimates or analogies with historical data. However, these techniques are often limited because they do not consider the existing knowledge of experts. In this paper, a novel cognitive data imputation technique is proposed for cost estimation that uses explanatory interactive machine learning to integrate and improve human knowledge. Through a case study in maintenance cost estimation the effectiveness of the approach is demonstrated.
© 2023 The Authors. Published by Elsevier Ltd on behalf of CIRP. This is an open access article under the CC BY license (http://creativecommons.org/licenses/by/4.0/)

1. Introduction and background

Cost estimation is a critical component of business success [1] as it provides an approximation of the cost outcomes of a programme, project, or operation. There are numerous challenges that are widely recognised in cost estimation, such as: lack of accuracy in estimates, insufficient availability of data, accounting for step changes with technology over time, lack of understanding of the product life cycle, use of unrealistic assumptions, and the dynamic nature of risks and uncertainties. Each of these factors is often tackled by the estimator incorporating his or her experience to arrive at an estimate with some degree of confidence. Whilst this adds to the level of uncertainty in the estimation process, it also often misses the opportunity to learn from good/bad inputs over time. It is also worth recognising that any estimator has only a limited amount of experience, however extensive that experience may be.

As a driving factor of inaccuracies in cost estimation, uncertainty can be considered as any deviation from the unachievable ideal of complete deterministic knowledge of the system in question [2]. Any dynamic system is subject to inherent uncertainty associated to it. This is due to both aleatory (e.g., irreducible, wider system driven), and epistemic uncertainty (e.g., lack of knowledge), which causes model-based predictions to differ from reality [3]. These categories of uncertainty are mostly experienced through data and models. Table 1 outlines the different sources of data uncertainty with corresponding examples.

The most common way to represent cost uncertainty is through probabilistic methods, in which uncertainty in data is represented by probability density functions, with triangular and normal being the most popular [4]. Furthermore, analytical and computational methods such as Monte Carlo simulation has widely been used with typically experts determining the uncertainty ranges based on experience [5]. Whilst probabilistic methods tend to be suitable for characterising aleatory uncertainty, this is not as common when there is a lack of statistical data where the uncertainty is caused by a lack of knowledge (epistemic uncertainty) [6]. This drawback has led to the investigation of the possibilistic approaches, which set the context for this paper. Possibility theory is a form of artificial intelligence (AI) [7] capable of representing uncertainty with much weaker statements of knowledge and more diverse types of data uncertainty [8]. However, machine learning (ML) based approaches are limited in their inferences to reasoning based on historical data [9].

Table 1
Sources of aleatory (A) and epistemic (E) data uncertainty.

Classification	Source	Type	Example
Variability	Inherent randomness	A	Metrics such as Repair Time, Mean Time Between Failure
Completeness	Lack of data	*E + A*	Data not entered by the operator.
Vagueness	Language	*E + A*	The component needs to be replaced *about* every 2 to 3 months.
Ambiguity	Multiple sources of data	E	Expert 1 and 2 provide different values to end of life costs.
Subjectivity	Optimism bias	E	Over confidence in schedule allocation.
Imprecision	Future decision or choice	E	Lack of understanding of whether Supplier A *or* B will be in business in the future.

* Corresponding author.
E-mail address: j.a.erkoyuncu@cranfield.ac.uk (J.A. Erkoyuncu).

https://doi.org/10.1016/j.cirp.2023.03.036
0007-8506/© 2023 The Authors. Published by Elsevier Ltd on behalf of CIRP. This is an open access article under the CC BY license (http://creativecommons.org/licenses/by/4.0/)

Extreme events such as pandemics and wars can completely overturn the assumptions used to train the AI models.

Accordingly, this paper aims to close a research gap in how current ML approaches do not fully exploit learning as a co-adaptive process, in which human' cognitive abilities are used to modify the behaviour of a ML system, but humans also adapt to use an ML tool more effectively and adjust their goals in response to what they learn by using the tool. The ML approach that can enable such a two-way collaborative approach is called explanatory interactive machine learning (XIL) [10]. In XIL, the ML-based system or the so-called learner interacts with the human user in a way that shows the user both its prediction as well as the explanation of why the predicted value is the way it is, and then the user can improve the learner's performance through his or her feedback [11]. The learner's feedback to the user can enhance the user's knowledge and the user's feedback to the learner can improve the learner's performance. Thus, this paper proposes an XIL-based two-way collaborative approach between humans and a ML-based system in the context of cost estimation. A case study in the maintenance domain is presented, showing that both ML imputation and estimation by a human can be improved through mutual interaction.

2. Cost estimation with explanatory interactive ML

To develop an XIL approach to cost estimation, we first define the relationships between the human expert, the ML system, and the data, as shown in Fig. 1.

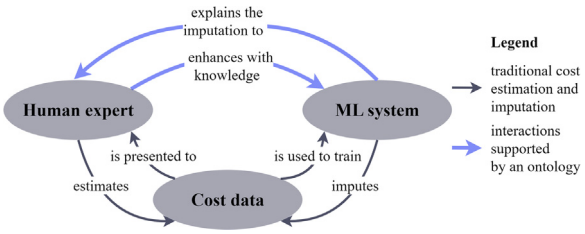


Fig. 1. Relationships between the human, the ML system, and the data. The interactions between the human and the ML system, which are supported by an ontology, emphasised.

The human typically interacts with the data through a graphical user interface (GUI) that presents the data in a readable form. The human can then use his or her skills as an expert to estimate costs and improve the data, e.g., by estimating missing values. In parallel, an ML system can be trained on the data and used to impute the missing values or predict future costs. The premise of XIL, however, is that human experts can improve the performance of the ML system by providing expert knowledge while at the same time improving their estimates through explanations provided by the ML system. The question is how this can be achieved in practise. In the presented approach, there are two main contributions to this: (1) to facilitate the interaction between the human and the ML system, an ontology-based approach is proposed, and (2) a specific ML method is chosen that is explainable and interactive.

The ontology shown in Fig. 2 contains the domain semantics (item, part, component, cost, etc.) defined according to the ISO standards and introduces additional entities and relations, *neighbour* and *is_similar_to*, which describe the relatedness or similarity between parts. An entity *Neighbour* is created due to the many to many cardinality of the *is_simlar_to* relationship. Similarity between components can be expressed in terms of certain criteria, such as geometry, material, functionality, etc. To provide a computational measure of similarity, the distance between the components $d_{i,j}$ is introduced and calculated, according to the selected criteria, for all pairs of components i and j, and gathered into a distance matrix $\boldsymbol{D}$. All distances are normalised. Hence, the data, the relationships between product, assemblies, parts, and components, the ML results and finally the expert knowledge are stored in a knowledge graph database using the ontology. The main advantages of using such a system include enabling imputation at different item levels (components, assembly or product level), and also taking into account items having similar function but different structure.

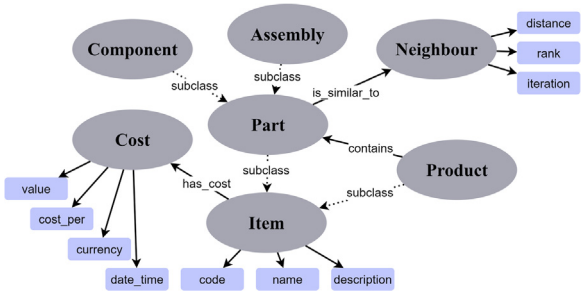


Fig. 2. Domain level ontology for cost estimation in maintenance.

2.1. Imputation and explanation methods

Imputation is performed using the k-nearest neighbours (kNN) algorithm. With kNN, the cost of an unknown component is calculated based on the similarity of the component to other components. The similarity is measured using a selected distance metric. Unlike black-box methods, kNN offers two main advantages: (1) interpretability and explainability, and (2) the ability to incorporate user feedback by modifying the distance metric. In general, the kNN method can be easily applied to cost estimation. The unknown cost is determined by the cost of the k most similar components, as shown in Eq. (1).

$$\hat{c}_i = \frac{\sum_{j=1}^{k} \frac{c_j}{d_{i,j}}}{\sum_{j=1}^{k} \frac{1}{d_{i,j}}} \tag{1}$$

The imputed cost, denoted by $\hat{c}_i$, is estimated using the actual costs of k neighbouring components, denoted by c_j, and the distances between them, denoted by $d_{i,j}$. To evaluate the ability of imputation, the predicted and actual costs are compared using the R^2 metric (Eq. (2)).

$$R^2 = 1 - \frac{\sum_{i=1}^{N} (c_i - \hat{c}_i)^2}{\sum_{i=1}^{N} (c_i - \bar{c})^2} \tag{2}$$

Here $\bar{c}$ is the average of the costs of all components. The value of R^2 ranges from 0 to 1, where 1 means that the imputed costs exactly match the real costs and 0 if the imputation model does not explain any of the variations in the imputed costs.

Accounting for user feedback is challenging for several reasons. First, the feedback may be of varying quality depending on whether the user is an expert. Additionally, the feedback needs to be provided in a way that can improve the imputation method. This is addressed as follows. The users provide cost estimates and lists of similar components. The extent to which a user's feedback is considered based on an assessment of the users' cost estimation abilities. The user is periodically tested by comparing their estimates for known component costs. A user-specific R^2 is computed and used to update the imputed cost, the *is_similar_to* relationship, and the distance matrix $\boldsymbol{D}$.

Different approaches can be used to update the costs and distances based on user feedback. However, we propose to treat the R^2 metric as a discount factor γ. For example, imputed costs are updated according to the user estimate u, weighted by γ , as shown in Eq. (3).

$$c' = u \cdot \gamma + c \cdot (1 - \gamma) \tag{3}$$

Since expert users have a higher R^2 (or γ), their estimates are weighted more heavily. This can be taken further to update the distances to neighbours via γ^2, e.g., $d'_{i,j} = d_{i,j} \cdot (1 \pm \gamma^2)$ ($\pm$ depending on the (dis)similarity of the components) and then taken even further to

update the distances to neighbours of neighbours via γ^3. This captures the idea that components that are similar to a particular component are also somewhat similar to each other, albeit with a discount factor. In this way, multiple distances and *is_similar_to* relations can be changed based on a single human feedback, reducing the amount of the feedback required to improve the distance metric.

The interaction between the human and the ML system is bidirectional. Once a cost is imputed using kNN, the user can view an explanation of the imputation. The explanation shows a list of similar components that were used for the imputation. In parallel, the user can check the comparison between the results of the imputation and the actual cost of the components with known costs (ground truth) and evaluate the quality of the imputation.

3. Case study

The industrial case provider is a leading provider of critical, complex engineering services. It provides a wide range of services across marine (e.g., ships and submarines) and aerospace equipment (e.g., helicopters) across Europe. As its business operation relies heavily on the use of high-value assets, asset management is a crucial activity for the company. To better manage the assets, a significant amount of information, including maintenance, operations, and financial data sets are collected and processed to improve reliability, availability, maintainability, and safety. However, critical shortcomings that prevent accurate analytics are related to the quality of the data. In fact, the data suffers from issues encompassing completeness (significant proportion of data with missing values), consistency (non-uniformization, and ambiguous data), and accuracy (presence of outliers, and lack of precision).

By improving data quality, the full potential of data sets can be exploited, leading to improved management of assets (e.g. cost-effective and time-saving benefits). A real data set is collected and employed in the present study, with the aim to improve the data completeness uncertainty.

3.1. Dataset description

The dataset is focused on purchase orders collected for the aerospace focused asset . This is with the intention to potentially maximise asset availability and optimise usage, balance spare parts in stock, and minimise operating cost. The dataset contains approximately 14 000 rows of purchase orders on different aerospace asset spare parts and around 70 attributes. The main attributes consist of the asset information (part name, descriptions, the product where the asset is part of, the quantity), the cost (cost_per, date, time, and currency, the part cost, exchange rate), and the storage information (warehouse-related).

A pre-analysis of the data revealed the presence of formatting issues, typos, redundant data, inaccurate part names, outliers, data out of range, and missing values of up to 20% in several attributes. In order to select the most important variables for cost imputation, Pearson correlation and Goodman tau were computed respectively for quantitative and qualitative variables of the dataset. From the correlation heatmap Cost_Per, Part_Description, Part, Currency, and purchase data were identified as critical variables for the analysis.

3.2. Imputation method analysis

In the present case, we use the product codes and the textual description of the components as the basis for evaluating the distance. The Levenshtein distance, which evaluates the number of atomic operations (insertion, deletion, replacement) required to transform one string into the other, is used as the distance metric. The distance between each pair of components is computed, normalised based on the maximum value, and collected in the matrices $\boldsymbol{D}_{code}$ for component codes and $\boldsymbol{D}_{description}$ for their descriptions. The textual description is not an adequate measure of similarity, but it can be used as a first approximation, to be improved later through user interaction. Fig. 3a) shows that the string similarities can sometimes predict the actual costs accurately (imputed costs close the diagonal),

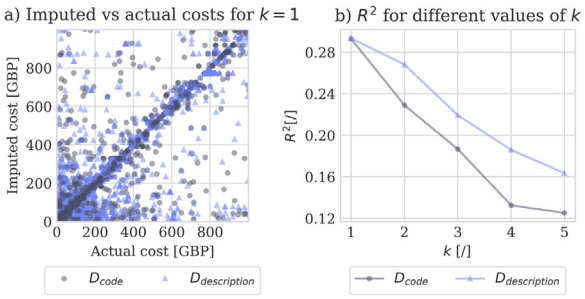


Fig. 3. An analysis of kNN imputation using the Levenshtein distance of part codes and part descriptions as the similarity metric.

but have low R^2 scores in general, dropping as the number of neighbours for imputation k is increased, as shown in Fig. 3b).

The question is how to achieve improvements by interacting with the user. To this end, first, a simulation study is conducted to validate the presented approach, and then the presented approach is also validated through a real-life experiment involving real human users, both experts and non-experts.

3.3. Simulation study

The study simulates interaction with artificially generated users. The main goals are to evaluate the possibility of improving the system through interaction and to estimate how much interaction is required for the improvement. The user model is as follows. User feedback consists of two components: determining an unknown cost for a given component and identifying similar components. For the cost determination, the user is simulated stochastically as shown in Eq. (4).

$$u(\pi,\sigma) = \begin{cases} \mathscr{U}(0,1000) & \text{if } \pi \leq \mathscr{U}(0,1) \\ \mathscr{N}(\mu,\sigma\mu) & \text{otherwise} \end{cases} \tag{4}$$

Here π represents the probability of the simulated user setting the cost randomly between 0 and 1000 using the uniform distribution $\mathscr{U}$. With the probability of $1-\pi$ the simulated users input a cost that is sampled from a normal distribution $\mathscr{N}$ around the actual cost, with the

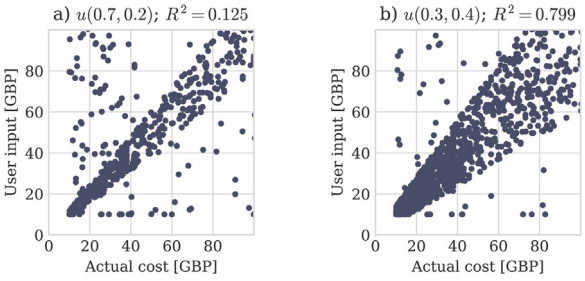


Fig. 4. Examples of simulated user inputs.

standard deviation σ expressed as the percentage of the actual cost. Fig. 4 shows simulated user inputs and the corresponding R^2 values.

The selection of similar components of simulated users is random and also based on their R^2. The higher the R^2, the more likely they are to choose a more similar component in terms of cost.

For the simulation, 100 components are selected at random and simulated to have missing cost values. In each iteration, a new, randomly

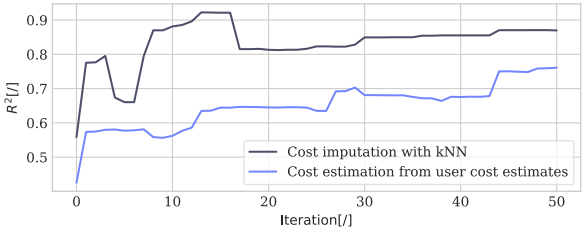


Fig. 5. A representative example of a simulation run showing improvement of cost imputation and estimation.

generated user is given 10 components for which they determine the cost and list similar components. The result is shown in Fig. 5.

Fig. 5 shows how the estimated component costs (blue) improve with subsequent user estimates (using Eq. (3)). The improvement is not monotonic, as simulated users can make large errors, which degrades the updated cost estimates. By having users list similar components and correct the distance matrix accordingly, the performance of kNN imputation is generally improved (dark grey). Thus, the simulation confirms that even stochastic users improve cost estimation and imputation using the presented methodology.

3.4. User study

A python web-based application was designed and implemented to enable human interaction and evaluation. The architecture is presented in Fig. 6. The GUI is developed in HTML5, jQuery 1.7.0, and Bootstrap 3.0, the backend is implemented in Django 2.0, and the

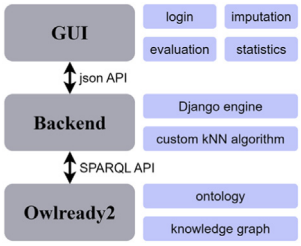


Fig. 6. The application architecture.

database in form of a knowledge graph using Owlready 2.0 that supports ontologies modelled using protégé.

The same methodology was used as in the simulation, except that instead of random guesses, real humans, both experts and non-experts, were involved in the estimation and imputation process. The users were also asked to identify components, similar in terms of cost. In order to minimise human effort, the users were assessed every five iterations of the imputation process. Fig. 7 shows the R^2 scores of 4 users, with users 1 and 2 having limited knowledge of the data and users 3 and 4 having advanced knowledge of the parts and their respective costs. The scores show an increasing trend over time,

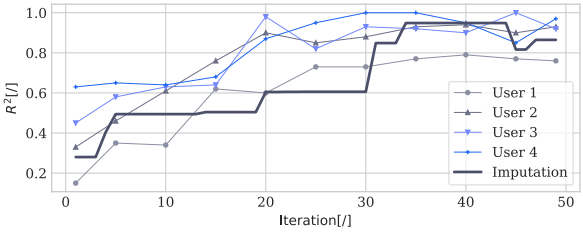


Fig. 7. Example cost estimation: user and algorithm improvement.

indicating an improvement in the estimates and the underlying human understanding of the data.

Imputation accuracy improves over time as well. The improvement shows a stepwise characteristic, similar to the simulation experiments in Fig. 5, due to the fact that a gradual update in the distance metric leads to discrete changes in nearest neighbours. In other words, as soon as a different set of neighbours is determined for a component, the performance can sharply increase or drop. A drop in performance of the imputation can be observed at iteration 44, indicating erroneous similarity assessment by the users. This implies that inappropriate or malicious users may degrade the kNN algorithm performance. Nevertheless, both the simulation study with random users as well as the user study show a robust growing trend in performance over time, confirming the feasibility of the presented approach.

4. Discussion and conclusion

This paper tackles data uncertainty related challenges in cost estimation. Whilst traditional cost estimation approaches have focused heavily on making use of existing data, they have typically missed the opportunity to benefit from expert cognitive human input to improve estimates. The approach proposed here enables to combine the expert knowledge with ML systems and offers two-way improvements leading to better estimates where there is data uncertainty. This is achieved by applying an XIL based approach supported by an ontology enabling the communication between human and ML method. We used a case study in aerospace maintenance with 14,000 lines of procurement related data to illustrate our approach through two stages, 1) a detailed simulation study, and 2) a user study. The results from both of these stages were complementary and confirmed the positive improvements in cost estimation from both user estimation and ML based imputation perspectives. For future research we aim to expand the data uncertainty challenges tackled, by moving beyond data gaps, to consider messy data scenarios. We also would like to evaluate the dynamic and evolving nature of data uncertainties for cost estimates, and explore the applicability of our approach to more complex problems.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

The research was supported by the EPSRC funded project DigiTOP (EP/R032718/1) and the Slovenian Research Agency (P2-0270). We also acknowledge Babcock International. Data supporting this study are openly available from (CORD) at DOI: 10.17862/cranfield.rd.22435462.

References

[1] Molcho G, Cristal A, Shpitalni M (2014) Part cost estimation at early design phase. *Annals of the CIRP* 63(1):153–156.
[2] Walker WE, Harremoes P, Rotmans J, van der Sluijs JP, van Asselt MBA, Janssen P, Krayer von Kraus MP (2003) Defining Uncertainty: a conceptual basis for uncertainty management in model based decision support. *Integrated Assessment* 4(1):5–17.
[3] Schuh G, Gützlaff A, Schollemann A (2022) Reduction of planning efforts for decision making under uncertainty in global production network design. *Annals of the CIRP* 71(1):385–388.
[4] Sakao T, Lindahl M (2015) A method to improve integrated product service offerings based on life cycle costing. *Annals of the CIRP* 64(1):33–36.
[5] Morse E, Dantan J-Y, Anwer N, Söderberg R, Moroni G, Qureshi A, Jiang X, Mathieu L (2018) Tolerancing: managing uncertainty from conceptual design to final product. *Annals of the CIRP* 67/2:695–717.
[6] Giovanni M, Stefano P, Tullio T (2011) Early cost estimation for tolerance verification. *Annals of the CIRP* 60(1):195–198.
[7] Dubois D, Prade H (2003) Fuzzy set and possibility theory based methods in artificial intelligence. *Artificial Intelligence* 148:1–9.
[8] Gila M, Asher C, Moshe S (2014) Part cost estimation at early design phase. *Annals of the CIRP* 63(1):153–156.
[9] Datta P, Roy R (2010) Cost modelling techniques for availability type service support contracts: a literature review and empirical study. *CIRP Journal of Manufacturing Science and Technology* 3(2):142–157.
[10] Rodriguez PEC, Guillebaud C, Wargnier H, Perry N (2022) Uncertainty and modelling cost based methodology for modelling choices in multiscale structures. *Annals of the CIRP* 71(1):129–132.
[11] Teso S, Kersting K (2019) Explanatory interactive machine learning. In: *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*, 239–245.



https://dspace.org/

Cognitive data imputation: case study in maintenance cost estimation

Erkoyuncu, John Ahmet; Namoano, Bernadin; Kozjek, Dominik; Vrabi, Rok

2023-01-01
Elsevier
https://doi.org/10.1016/j.cirp.2023.03.036
Erkoyuncu JA, Namoano B, Kozjek D, Vrabic R. (2023) Cognitive data imputation: case study in maintenance cost estimation. CIRP Annals - Manufacturing Technology, Volume 72, Issue 1, July 2023, pp. 385-388

Downloaded from a DSpace repository. https://dspace.org/