Available online at www.sciencedirect.com

ScienceDirect

Procedia CIRP 135 (2025) 888–893

www.elsevier.com/locate/procedia

32nd CIRP Conference on Life Cycle Engineering (LCE 2025)

Application of Matrix Completion Techniques on LCA Data for Different Recycling Scenarios for Parts of Professional Data Centers

Lisa Dawel[a]*, Felix Schmedes[a], Fernando Andres Penaherrera Vaca[a], Alexandra Pehlken[a]

[a]*OFFIS - Institute for Technology, Escherweg 2, 26121 Oldenburg, Germany, Country*

* Corresponding author. Tel.: +49 441 97 22 745; *E-mail address:* lisa.dawel@offis.de

Abstract

When modelling a product by using the LCA methodology, some gaps of knowledge need to be filled in. One solution is to use estimates from available data that could be outdated or only fit roughly the same category. Another solution is to use AI, which generalizes knowledge and can thus provide better estimates. LCA data has a unique structure that certain machine learning algorithms can use to their advantage like the linear dependency between a subset of metrics and scenarios. The research question is how good this generalization works on LCA datasets of limited size with their unique properties. It includes an analysis, which pre-processing techniques that are taking into account the unique structure of LCA data, can improve the prediction performance. Furthermore, the threshold the percentage of missing entries can reach while still ensuring a reasonable performance, is analysed. This paper is a case study that investigates the potential of matrix completion algorithms on LCA data on a small scale using recycling scenarios for parts of professional data centers to derive knowledge for bigger scales.

Keywords: LCA; Matrix Completion; Recycling; Professional Data Center; Case Study; Knowledge Generalization

1. Introduction

In our modern world, we need more sustainable solution in all areas of modern life. One important sector regarding green house gas emission is Manufacturing / Construction that accounted to approximately 13% of all world CO_2 equivalents emissions in 2021 [1]. This paper focusses on the manufacturing sector. It has a big influence on the everyday life of every human as products surround us all. The design of products has a big influence on the manufacturing sector. If products are designed more sustainably, the manufacturing emissions should decrease as well. Furthermore, if products are designed in such a way that they are repairable and/or last longer, the emissions from the manufacturing sector should decrease as well – at least in the long run.

The sources for this knowledge and all the sustainability KPIs is the Life Cycle Assessment (LCA), which models the whole production process of the product. However, in order to model the whole process a lot of reference data is necessary. This is very important to ensure that the model fits the application and to ensure that the results are realistic. The challenge arises that not all data is available in the Ecoinvent database or that the data available is incomplete. Hence, models have to be built on available data and might not be accurate. This paper aims to help mitigate this challenge by studying how good matrix completion algorithms work for LCA data. This is relevant due to the unique structure of LCA data. The research questions are defined as follows:

RQ1: How well do different matrix completion algorithms work for the unique structure of LCA data?

RQ2: Where is the threshold for missing entries so that the results are still reasonably good?

To study these research questions, a data set of professional data centers [2] was selected. The dataset is described in section 3.

2. Literature Review

Matrix completion is a well-known area of research for quite a long time [3]. Next to classical approaches that use properties

10.1016/j.procir.2024.12.088

of matrices to fill in missing information the Netflix problem brought new algorithms to fruition and sparked the use of deep learning algorithms in this field. Using matrix completion in the LCA promises many advantages. When modelling a product, not all data needed is always available. Then, data from the literature needs to be used to fill in the gaps, which is a time consuming process. However, the data that is available in databases like Ecoinvent, might not fit the application very well. For this, the matrix completion algorithm could be used to fill in the gaps in the data.

A research gap was identified in the area of completing missing data in life cycle assessments (LCA). This gap is particularly pronounced for LCA data that relate to server components. The majority of the studies analyzed focus on completing missing data within the Ecoinvent database, which plays a central role in conducting life cycle assessments [4]. To the best of our knowledge, only the following papers exist in this area.

In their paper, Fangfang et al. [5] present two approaches for estimating missing information in LCA and input-output analysis. Both approaches are based on the assumption that the data used in the analysis have a low-rank or almost low-rank structure due to their similar data structures [6] and input ratios. Based on this assumption, the authors used matrix completion techniques to reconstruct the missing data. Two non-negative matrix completion models were presented, which are based on the use of Alternating Direction Method of Multipliers (ADMM). The results indicate the usefulness of the methods used on the data that represent different production processes and their input-output relationships. Cai [7] focused in his work on the completion of LCI data, which also come from the Ecoinvent database. Data was used that contains information on various environmental inputs and outputs of industrial processes, such as material and energy consumption as well as emissions linked to the processes. To complete missing data, a similarity-based link prediction procedure was used, among other things. In addition, the author estimated missing values based on weighted similarity metrics. Canals et al. [8] used different approaches to fulfil gaps in LCA Data of bio-based products. In their paper they described the use of proxy data, a method where similar data of a product is used to estimate another. Also averaged proxies where the average of several similar products or scaled data where used to fill the gaps. Furthermore, data was extrapolated by adapting existing datasets to new products by changing parameters, such as production methods or regional characteristics. Zhang et al. [6] applied methods like Singular Value Thresholding (SVT) and Factor Group-Sparce Regularization (FGSR) to complete data in realtime. The examined data originates from a specific part of a distillation unit of a refinery. Imputed values were based on values of temperature, pressure and flow rate data. Results are showing a superior accuracy compared to traditional MC methods. Saad et al. [9] used an approach based on decision trees to close gaps in LCI data of the Ecoinvent database. The paper is focusing on GHG emissions in the manufacturing process of products. Models like Gradient Boosting showed high accuracy in prediction emissions.

Summarizing, in some of the papers presented above [6, 7, 9], the error is still remains relatively substantial, especially for small datasets. This makes these algorithms useful when the alternative is not having any data. However, the results have to be checked for plausibility in each case, which makes the process tedious. One possible reason could be outliers in the data that are not treated. The usage of robust pre-processing methods or robust matrix completion algorithms can be a solution. However, we found very few papers using robust matrix completion methods in LCA data to deal with impulsive noise or outliers like Wen et al. [10] or Fangfang et al. [5]

In this paper, we aim to close this research gap by analyzing matrix completion algorithms and deep learning algorithms with robust preprocessing for the specific field of datacenters.

3. TEMPRO Dataset

The data used for this is the dataset is the results of the project TEMPRO: Total Energy Management for Professional Data Centers". This project analyzed the components of professional data centers to construct updated models for LCA. Additionally, recycling scenarios for different key material

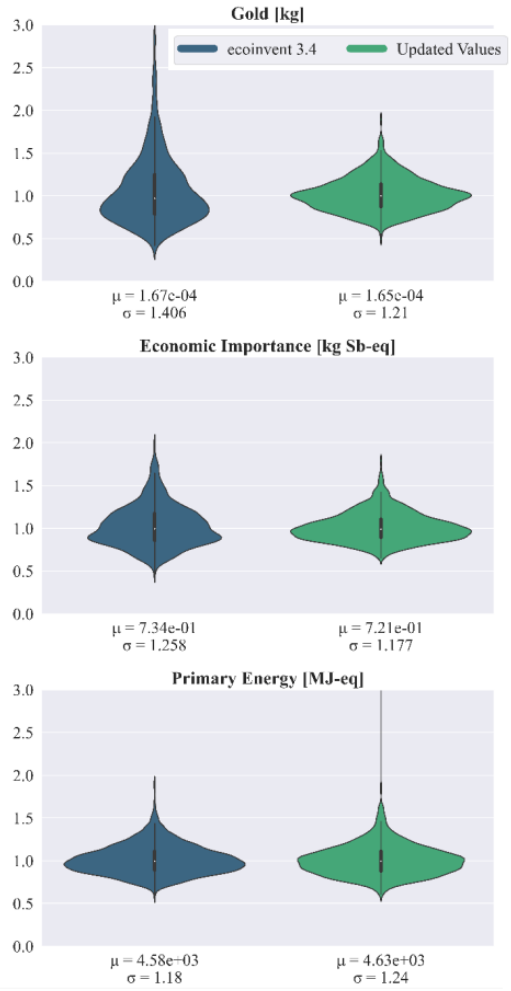

Figure 1: Comparison of result uncertainty values (after laboratory measurements) for PCBs. Literature data from ecoinvent 3.4 data was compared with own lab analyses of PCB samples (Source: PhD Dissertation Penaherrera, 2024)

were developed. Data on PCB components was extracted and analyzed. This ensures that the dataset is more accurate than data from literature and is highlighted in Figure .

The analysis of inventories at different granularity levels allows Material Flow Analysis to be conducted for the stages of the lifecycle. In addition, the use of Monte Carlo analysis paired with data quality and inventory uncertainty evaluation allows evaluating the quality of the results, and to evaluate the improvements on said quality obtained. A complete separate database with products built only based on reference ecological databases was constructed in Penaherera [2] to quantify the improvements achieved through the incorporation of data from laboratory analysis. While a clear improvement can be observed in the material-related indicators at the resource accounting. This reflects the methodologies for construction of such indicators and shows the effects of incorporating more material flows and higher quantities of material use for manufacturing of electronic components. Reporting of the uncertainty increases in general the quality of any LCA study, with a full quantitative uncertainty assessment giving high transparency on the quality of the results provided.

By using a variety of indicators for evaluation, this work presents novel insights on material depletion and on energy use for various products, whose high specific critical material content make them of relevance for urban mining and for securing secondary material sources. Additionally, the use of Monte Carlo simulations paired with data quality to evaluate the results is a procedure not commonly observed within LCA studies, mostly due to time and to computer resource constraints. This last evaluation provided an insight on methodologies for evaluation of results quality and on the improvement of said quality, while also providing a basis for studying interdependences of indicators. However, limitations on the methodology are reflected when evaluating the distribution of the results. Therefore, this research will be continued in combination with the matrix completion.

The modeling of recycling and the evaluation of the results of each scenario was done using LCA of the product, material flow analysis, and evaluation of recycling processes. Recycling is modeled in two main steps: Pretreatment and metal recovery.

Many metals are concentrated on certain parts of the WEEE (Waste from Electrical and Electronic Equipment) components. Pretreatment has the goal of separating different portions with concentration of a specific metal or metal family. Disassembly of these parts is the most time-consuming operation. Automatic, semiautomatic, and manual disassembly systems have been developed, the latter being the most adopted technique. The recovery efficiency by manual treatment is a lot higher than that of automatic systems. Manual sorting and dismantling are economically unfeasible in developed economies.

Several pretreatment processes where considered to develop models and scenarios, where the main difference lies on the amount of material loss in each pretreatment chain.

Pretreatment processes considered are manual sorting and dismantling, multilevel deep manual dismantling (which includes manual separation of soldered components such as integrated circuits), mechanical/automated sorting and separation, or a combination of both.

After pretreatment, four material recovery processes were evaluated. Pyrometalurgical recovery, the most widely used process in industry, involves the use of elevated temperature process to extract metals. Hydrometallurgical recycling uses techniques to leach metals into solutions during reactions with leachants and oxidants. Products are afterward separated and purified. Electrochemical recovery involves separation of base metals and precious metal containing fractions. It has the advantage of a lower use of chemical agents. Biometallurgical processes use special microbes for metal extraction, with an emphasis on recovery of copper and gold.

The part of the dataset that is used contains multiple levels: different data centers, their systems (e.g. climatization, server room, server rack), their devices (e.g. server 1U, storage or power distribution unit), their different parts (e.g. CPU, HDD, Mainboard). Moreover, models for the production of their constituent materials were developed. Therefore, there are four levels of data in this dataset. Data centers contain very different parts and outliers are present. We are investigating the influence of these outliers in section 5. For the whole dataset, 172 different components are used.

From the models, different impacts were calculated using different methodologies [2]. From this, the following six LCA KPIs (key performance indicators) have been selected: Greenhouse gas emissions [kgCO2e], ReCiPe-Total Impacts, Geo-Political Supply Risk, ADP – Economic Importance [kg Sb-eq], Primary Exergy Demand [MJ-eq], and Primary Energy Demand [MJ-eq].

Thus, a four-dimensional dataset, a data cube, is created with the four dimensions as follows: (1) One dimension contains the components of the datacenters with the levels described above. (2) One dimension captures the recycling scenarios, (3) the next the sorting scenarios. (4) The last dimension contains the six LCA KPIs.

The term 'dimensionality' refers here to the structural arrangement of the data in terms of axes within the matrix. In dimensionality reduction, it however refers to reducing the number of columns in a two-‘dimensional’ matrix. Mitigating this ambiguity of the term ‘dimensionality’, we now use the term ‘tensor’. The matrix described above is a tensor of rank four, with the dimensions (172, 6, 5, 4) meaning that is has 172 entries with 6 LCA KPIs, 5 sorting categories and 4 recycling categories each. Since most machine learning algorithms are not designed to work with data cubes with a rank higher than two, the matrix is flattened into a vertical form with rank two.

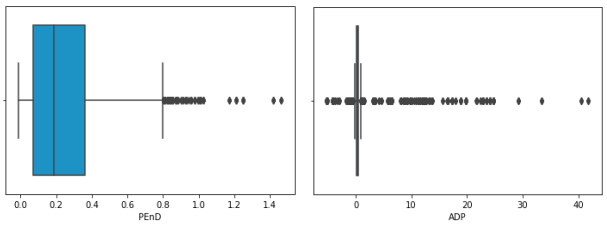

Figure 2: Boxplots of two different LCA KPIs. As it can be seen, outliers are present for some KPIs. Also, it can be seen, that the distribution is not symmetrical.

Before, the index is the ID (identification number) of the product or component, now the index is the combination of ID, sorting category and recycling category. In the columns, the LCA KPIs are shown.

This structure is universal for LCA data as anticipated in this paper because different product design scenarios can replace the combination of recycling and sorting scenarios, as they are equivalent from a data structure point of view.

Because the columns, i.e. the LCA KPIs, are not linearly dependent, the rank is always the number of columns n due to

$$rank(A) \leq min(m, n),$$

when A is a $(m\ x\ n)$ matrix or, equivalently, a tensor with rank 2 and dimensions(m, n). m is larger than n as it represents the number of components with all their scenarios, respectively.

4. Method

4.1. Preprocessing

4.1.1. Normalization and Standardization

Two different methods for data normalization were used.

The standardization (s) with mean and variance. This method centers the data on the mean and scales it to unit standard deviation. This method is highly influenced by the presence of outliers because it assumes a Gaussian distribution. [11].

The robust (r) standardization with median and interquartile range (IQR). This method is robust in the presence of outliers.

4.1.2. Dealing with outliers

Identifying and handling outliers is important, as they can significantly affect the matrix completion process. In the data set used, outliers present, as can be seen in Figure 2. Next to the robust scaling with median and IQR as described in section 4.1.1, winsorization (w) is used. This method limits extreme values to reduce the effect of outliers. It has the disadvantage of having a hard border, which labels data points as outliers or not. When using winsorization, a trade-off has to be made between limiting the influence of outliers and changing the distribution. In our case, the level of winsorization was customized to each LCA KPI due to their different distributions.

4.2. Matrix Completion Algorithms

4.2.1. Statistical Matrix Completion Algorithms

Using the software package fancyimpute [12], the following statistical matrix completion algorithms (SMCA) are used:

As a baseline, the algorithms Mean Fill and Median Fill which fill the gaps with the mean or the median, respectively.

One of the easiest matrix completion algorithms is the k-nearest neighbor (KNN) algorithm [13] with k being the number of nearest neighbors that are used to fill in gaps in the data using the mean squared difference for the features of data points that both have data present.

Softimpute is an algorithm, which completes matrices by iterative soft thresholding of SVD decompositions [14].

Finally, the Iterative Imputer is used which models each feature with missing values as a function of other features in a round-robin fashion [15, 16]. All the algorithms that are used can only work with the vertical form of the data set – as described in section 3. Furthermore, they can only work with numerical data, not with categorical. These algorithms have, however, the advantage of a fast computing time and their ability to work with small data sets. Since they are applied directly to the whole dataset, there is no difference between training and inference; there is no need for a training test split.

4.2.2. Matrix Completion with Auto Encoders

Another class of matrix completion algorithms are auto encoders (AE), which have gained importance in the field of recommender systems and collaborative filtering to fill gaps of missing data. In particular denoising auto encoders (DAE) achieved great results in the task of imputing missing data with a high accuracy. AE in general have the ability to model complex and non-linear patterns in data. DAE aim to reconstruct the exact input data, which can be beneficial for scenarios where the goal is to recover the original values as accurately as possible [17].

For the reconstruction of missing LCA Data we used a DAE in the following way. Our denoising autoencoder was designed to learn efficient representations of data while reducing noise. It corrupts the input data with Gaussian noise, compresses it into a lower-dimensional latent space using a ReLU-activated encoder, and reconstructs the data via a sigmoid-activated decoder. The noise factor was set to 0.2. The model is trained using the Adam optimizer and mean squared error (MSE) loss.

We trained the DAE for 50 epochs and implemented an early stopping mechanism to halt training if there was no improvement for five consecutive epochs. The architecture of the model comprises an input layer that accepts six inputs (for

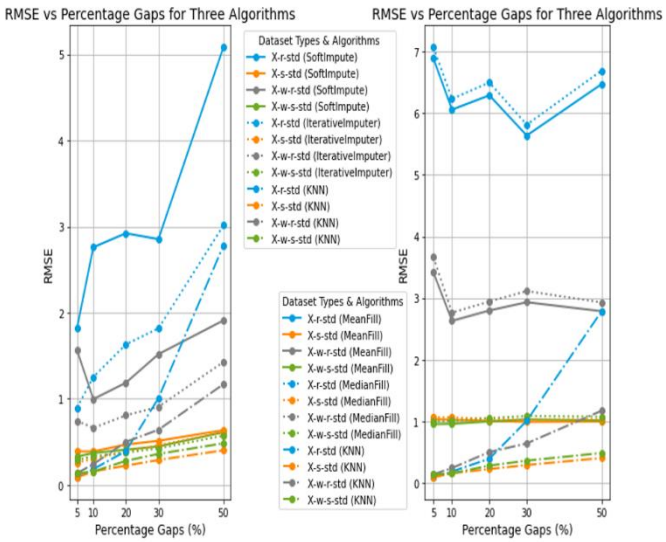

Figure 3: Comparison of different algorithms described in section 4.2.1 for different pre-processing methods described in section 4.1. On the x-axis the amount of artificial gaps introduced to the data set are shown. On the y-axis the error value for each algorithm is depicted. r_std stands for robust preprocessing, s_std for standardization with mean and standard deviation and w stands for winsorization as described in section 4.1.1 and 4.1.2

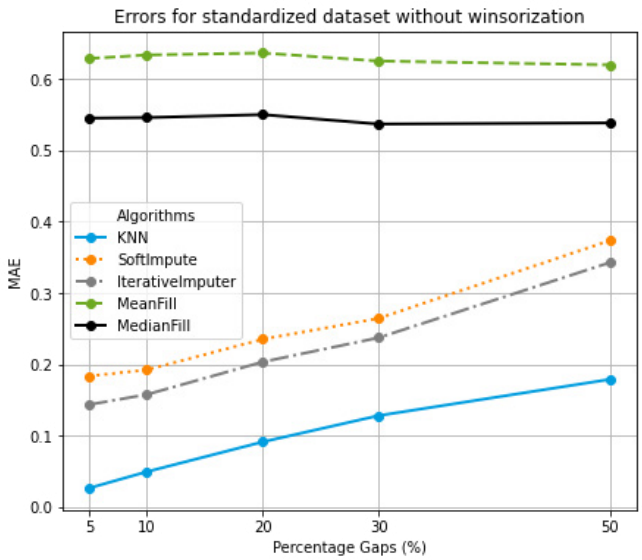

Figure 4: Comparison of the different algorithms described in section 4.2.1 for one pre-processing method: standardization with mean and standard deviation. The KNN works best, SoftImpute and IterativeImpute have a similar performance. As expected have Mean and Median Fill the worst performance as they are for reference only. Median Fill works better than Mean Fill which could be due to the outliers in the data. On the x-axis the number of artificial gaps introduced to the data set are shown. On the y-axis the error value for each algorithm is depicted. As expected, the error rises with increasing percentage gaps for most algorithms.

each scenario the six results of each metric), followed by a dense layer with three neurons, and another dense layer that reconstructs the data back to its original dimensionality of six.

To improve the robustness of our model, Gaussian noise was added in advance to a specific fraction of the data before training (10%, 20%, 30%, and 40%). The data was scaled using a MinMaxScaler to normalize the values to a range between 0 and 1. Due to the limited size of the dataset for deep learning, the data is split in a train-test-validation ratio of 80-10-10.

5. Results

To assess the results of the applied algorithms we are using three different metrics: the root mean squared error (RMSE), the mean squared error (MSE) and the mean absolute error (MAE).

5.1. Statistical Matrix Completion Algorithms

The algorithms described in 4.2.1 are applied to the dataset that is pre-processed as described in section 4.1. The results are shown in Figure 3 and Figure 4. It is obvious, that KNN is the best performing algorithm followed by SoftImpute and IterativeImpute that perform similarly. The performance of the benchmarking mean and median fill are the worst. As seen in 4, the error of KNN with $k=3$, is $1/10^{th}$ of the error of mean and median fill.

The results show, that for this particular data set, outlier treatment is unnecessary: In Figure 3, it is evident that the winsorization increases the error. In addition, the robust scaling does not bring any advantages. This means that these pre-processing steps can be left out. Standardization is however very important for most algorithms.

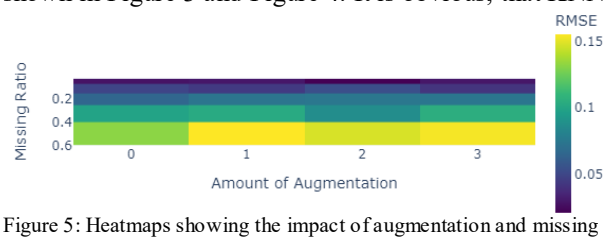

Figure 5: Heatmaps showing the impact of augmentation and missing data on error metrics (MSE, RMSE, MAE).

5.2. Auto Encoder

Table 1. Auto Encoder results

Amount augmentation	Missing ratio	MSE	RMSE	MAE
1	0.05	0.0004	0.0208	0.0035
2	0.05	0.0006	0.0230	0.0038
3	0.05	0.0007	0.0252	0.0045
0	0.05	0.0008	0.0268	0.0046
0	0.10	0.0012	0.0318	0.0076
2	0.10	0.0012	0.0325	0.0076
1	0.10	0.0019	0.0399	0.0094
3	0.10	0.0035	0.0468	0.0121
3	0.20	0.0034	0.0563	0.0178
2	0.20	0.0040	0.0624	0.0185
1	0.20	0.0079	0.0763	0.0258
0	0.20	0.0103	0.0828	0.0286
1	0.30	0.0082	0.0866	0.0330
0	0.30	0.0142	0.1004	0.0415
3	0.30	0.0172	0.1061	0.0440
2	0.30	0.0175	0.1061	0.0454
3	0.50	0.0220	0.1309	0.0696
0	0.50	0.0251	0.1439	0.0719
2	0.50	0.0315	0.1492	0.0774
1	0.50	0.0332	0.1511	0.0792

Table 1 shows the performance of matrix completion using an auto encoder, measured by MSE, RMSE and MAE. As the missing ratio increases, error metrics generally rise, indicating that higher missing data leads to poorer imputation performance. Lower missing ratios and augmentation levels result in better accuracy, demonstrating the model's limitations with increased data sparsity.

Figure 5 shows a heatmap of the results of the AE models. The graph shows that an increase in missing data points is accompanied by an increase in the error metric. In general, moderate augmentation (around 1 or 2) seems to help reduce RMSE, particularly for higher missing ratios (e.g. missing ratio 0.20 with augmentation level 3). Too much augmentation or no augmentation at all often results in worse performance at higher missing ratios.

6. Conclusion and Future Work

To summarize, the paper shows the performance of the different algorithms and preprocessing methods for matrix completion. Ultimately, if the size of the data set is large enough – as in this case - DAE can work very well with LCA data. The performance of the SMCAs is comparable to the performance of the AE. Our results suggest a high influence of the pre-processing steps for the SMCAs Depending on error type and algorithms, most algorithms performed well until up to 30% missing values rendering it very useful. When applying matrix completion methods to other data sets, it is necessary to select preprocessing methods and algorithms fit the unique properties of the data and the needs of the application.

DAEs have the advantage, that once trained they can be used for multiple inferences even without the training data set present. However, the training data set needs to be a full data set without any gaps. SMCAs can work with much less data and can impute gaps directly in the data set.

In future work we want to try out other pre-processing methods with further matrix completion algorithms. In addition, the combination of physical models and methods of artificial intelligence can further improve the models, especially when data is scarce. This could be achieved with models like [18, 19]. Furthermore, we want to apply our knowledge to other data sets to see if the results are replicable.

Acknowledgements

Funded by the European Union. Views and opinions expressed are, however, those of the author(s) only and do not necessarily reflect those of the European Union or the European Health and Digital Executive Agency (HADEA). Neither the European Union nor the European Health and Digital Executive Agency (HADEA) can be held responsible for them. Funding NO101091490.

References

1. Climatewatch data Greenhousegas Emissions from Manufacturing / Construction Sector. https://www.climatewatchdata.org/data-explorer/historical-emissions?historical-emissions-data-sources=climate-watch&historical-emissions-gases=&historical-emissions-regions=&historical-emissions-sectors=All%20Selected&page=1. Accessed 25 Sep 2024
2. Vaca FAP (2024) Analysis of Interactions Between Raw Material and Energy Demands for Data Centers
3. Johnson CR (1990) Matrix theory and applications: [lecture notes prepared for the American Mathematical Society Short Course Matrix Theory and Applications, held in Phoenix, Arizona, January 10-11, 1989]. Symposium in Applied Mathematics <American Mathematical Society>: Proceedings of Symposia in Applied Mathematics, 40 : AMS short course lecture notes. American Math. Soc, Providence, RI
4. Ecoinvent Introduction to the Database. (https://support.ecoinvent.org/introduction-to-the-database). Accessed 25 Sep 2024
5. Fangfang Xu, Chen Lin, Guoping He, Zaiwen Wen Nonnegative Matrix Completion for Life-cycle Assessment and Input-ouput Analysis
6. Zhang X, Sun X, Xia L et al. (2024) A Matrix Completion Method for Imputing Missing Values of Process Data. Processes 12:659. https://doi.org/10.3390/pr12040659
7. Jiarui Cai Computational Approaches for Estimating Life Cycle Inventory Data
8. Canals LMi, Azapagic A, Doka G et al. (2011) Approaches for Addressing Life Cycle Assessment Data Gaps for Bio-based Products. Journal of Industrial Ecology 15:707–725. https://doi.org/10.1111/j.1530-9290.2011.00369.x
9. Saad M, Zhang Y, Jia J et al. (2024) Decision tree-based approach to extrapolate life cycle inventory data of manufacturing processes. Journal of Environmental Management 360:121152. https://doi.org/10.1016/j.jenvman.2024.121152
10. Wen S, Xu F, Wen Z et al. (2014) Robust linear optimization under matrix completion. Sci China Math 57:699–710. https://doi.org/10.1007/s11425-013-4697-7
11. Zoubir AM, Koivunen V, Ollila E et al. (2018) Robust statistics for signal processing. Cambridge University Press, New York, NY, USA
12. Rubinsteyn A, Feldman S fancyimpute: An Imputation Library for Python. https://github.com/iskandr/fancyimpute
13. Hastie, T., Tibshirani, R., Friedman, J. H., & Friedman, J. H. (2009) The elements of statistical learning: data mining, inference, and prediction
14. (2010) Spectral regularization algorithms for learning large incomplete matrices
15. (2000) Multivariate imputation by chained equations
16. Buck SF (1960) A Method of Estimation of Missing Values in Multivariate Data Suitable for Use with an Electronic Computer. Journal of the Royal Statistical Society Series B: Statistical Methodology 22:302–306. https://doi.org/10.1111/j.2517-6161.1960.tb00375.x
17. Cardoso Pereira R, Seoane Santos M, Pereira Rodrigues P et al. (2020) Reviewing Autoencoders for Missing Data Imputation: Technical Trends, Applications and Outcomes. jair 69:1255–1285. https://doi.org/10.1613/jair.1.12312
18. Reuter MA, van Schaik A (2015) Product-Centric Simulation-Based Design for Recycling: Case of LED Lamp Recycling. J Sustain Metall 1:4–28. https://doi.org/10.1007/s40831-014-0006-0
19. Reuter MA, van Schaik A, Gediga J (2015) Simulation-based design for resource efficiency of metal production and recycling systems: Cases - copper production and recycling, e-waste (LED lamps) and nickel pig iron. Int J Life Cycle Assess 20:671–693. https://doi.org/10.1007/s11367-015-0860-4