Machine learning for gap-filling in greenhouse gas emissions databases

Luke Cullen
Department of Engineering
University of Cambridge
Cambridge, UK
lshc3@cam.ac.uk

Andrea Marinoni
Department of Physics and Technology
UiT the Arctic University of Norway
Tromsø, Norway
andrea.marinoni@uit.no

Jonathan Cullen
Department of Engineering
University of Cambridge
Cambridge, UK
jmc99@cam.ac.uk

Abstract

Greenhouse Gas (GHG) emissions datasets are often incomplete due to inconsistent reporting and poor transparency. Filling the gaps in these datasets allows for more accurate targeting of strategies to accelerate the reduction of GHG emissions. This study evaluates the potential of machine learning methods to automate the completion of GHG datasets. We use 3 datasets of increasing complexity with 18 different gap-filling methods and provide a guide to which methods are useful in which circumstances. If few dataset features are available, or the gap consists only of a missing time step in a record, then simple interpolation is often the most accurate method and complex models should be avoided. However, if more features are available and the gap involves non-reporting emitters, then machine learning methods can be more accurate than simple extrapolation. Furthermore, the secondary output of feature importance from complex models allows for data collection prioritisation to accelerate the improvement of datasets. Graph based methods are particularly scalable due to the ease of updating predictions given new data and incorporating multimodal data sources. This study can serve as a guide to the community upon which to base ever more integrated frameworks for automated detailed GHG emissions estimations, and implementation guidance is available at https://hackmd.io/@luke-scot/ML-for-GHG-database-completion.

1 Introduction

Greenhouse Gas (GHG) emissions datasets are often incomplete both at facility level (1) and at national level (2). Countries and companies are accelerating emissions reduction strategies inline with the Paris agreement and net-zero objectives (3; 4; 5; 6), but incomplete and inaccurate datasets remain a barrier to effective policy-making (7; 8; 9). Large companies are required to use emissions intensity factors, provided by government entities or taken from life-cycle assessment databases (10; 11; 12), to convert their facilities' activity data to GHG emissions. The 3 main causes of incompleteness at this level are: companies excluded from reporting regulations, lack of transparency, and non-compliance (9; 13). For companies, the resulting uncertainty, can lead to erroneous net-zero calculations and missed emissions (14). The facility-level data is then aggregated to a national estimate grouped in to source types and reported yearly by UNFCCC annex I parties. Only 43 out of 198 countries are considered annex I, accounting for $\sim$20% of global emissions, while non-Annex I countries have

Tackling Climate Change with Machine Learning workshop at NeurIPS 2023

inconsistent reporting (15). National reports are further limited by the standardised breakdown used by the UNFCCC and the accumulation of these uncertainties can lead to the misunderstanding of emissions reduction progress and misinformed policy decisions (15).

The most accurate gap-filling solution is to require all emitters to conduct full life-cycle assessments for their products, but this is not viable at large scale due to the expense and difficulty involved (16; 17). Data fusion, combining multiple data sources has resulted in some country-wide and industry-wide emissions databases (18; 19; 8; 20; 21), but these either lack facility-level granularity or have incomplete coverage of facilities. A second option, which we will explore in this study, is to improve data coverage and quality by 'gap-filling' emissions datasets using the available data to infer the unavailable data. Inference and machine learning methods have been implemented for improving process emission estimates in life-cycle assessments (22; 23; 24; 25), quantifying uncertainty in material flow analysis (26; 27), and data mining for industrial ecology (28; 29), but have so far been unsuccessful in capturing the complexity and types of inputs required for full GHG emissions estimates (30; 31; 32). Gap-filling emissions datasets is a crucial problem where data science may help if industrial ecologists are equipped with the knowledge of which machine learning techniques can be used to tackle which problems. This paper seeks to address this gap and respond to the call for data-driven innovation in the community (32; 9; 33) by introducing 18 classification models that can be applied to the 3 types of gap-filling problem that we define in section 2. Section 3 will assess the performance of each model when applied to gap-filling problems across 3 GHG emissions databases and section 4 will discuss the implications of the results. An interactive guide is available at `https://hackmd.io/@luke-scot/ML-for-GHG-database-completion`. It can serve as a reference for the community and provide a base for future development of multimodal frameworks for inferring company-specific emissions estimations, inline with urgent needs from companies and governments as they attempt to achieve their net-zero targets.

2 Methods

In table 1 we present 3 types of gap addressing increasingly challenging problems in GHG emissions datasets. Level 1 considers an emitter with some reporting records, but with missing time steps. Level 2 considers an emitter with no reports at any time step, but with properties that are shared with other emitters in the database. Level 3 considers an emitter with no reports at any time step and with at least one property that is not shared with any other emitter in the database, for example for a facility in a country with no reporting facilities at all.

Table 1: Levels of gap filling. E is the set of emitters and P is the set of properties of those emitters.

Level	Gap composition	Set relationship	Inference type	Typical use case
1	Time step	$E_{test} \subseteq E_{train}$ $P_{test} \subseteq P_{train}$	Supervised learning	· Inconsistent data reporting from well regulated facilities.
2	Emitter	$E_{test} \nsubseteq E_{train}$ $P_{test} \subseteq P_{train}$	1-D transfer learning	· Non-reporting small facilities. · Poor transparency.
3	Emitter with unknowns	$E_{test} \nsubseteq E_{train}$ $P_{test} \nsubseteq P_{train}$	Multidimensional transfer learning	· UNFCCC gaps, especially non Annex-I countries. · Poorly mapped products.

The 3 levels of gap-filling will be considered across 3 datasets which cover a range of typical properties of GHG emissions databases: UNFCCC (2) - national-level with 2 features, ClimateTRACE (1) - facility-level with 6 features, Petrochemical - facility-level with 13 features (see SI1.1 for feature details). We will discretise gap-filling to a 4-class classification problem with class boundaries set according to the quantiles of emissions values in each dataset as follows: $0 \rightarrow 0.25$ - low emissions, $0.25 \rightarrow 0.5$ - medium emissions, $0.5 \rightarrow 0.75$ high emissions, $0.75 \rightarrow 1$ very high emissions. For each gap level, considered for each dataset, we will apply a series of classification techniques and evaluate their performance using average accuracy and $F1$ score. The models used are assembled into 5 types: interpolation (34), shallow learning models (35), ensemble models (36), deep learning models (37) and graph representation learning models (38). Model structure details can be found in SI1.2 and cross-validation is used to optimise hyperparameters and prevent overfitting. Finally, feature importance will be used to establish the most valuable features to collect during future data gathering and will be evaluated for each model.

3 Results

Figure 1 displays the average accuracy for each model applied to each dataset and for each level of gap-filling, given a constant 70%-30% train/test set split. For level 1, with missing years in the UNFCCC dataset, interpolation is the best solution with 97% average accuracy. In datasets with more input features, decision trees and nearest neighbours perform as well or better than interpolation at level 1. Deep learning models are not well suited to optimally predict values for a single emitter and have lower accuracy than as some simpler methods. For level 2 with missing emitters, interpolation is not the best option for any dataset. Decision trees and random forest methods are viable solutions across all datasets with 60%-70% accuracy. For the UNFCCC dataset with only two input features, deep learning is ineffective, but for datasets with more input features, deep learning and graph representation learning methods perform as well as the best shallow methods. For level 3, interpolation and shallow models have poor accuracy. For the ClimateTRACE and petrochemical datasets, deep learning and graph-based models provide the best solutions.

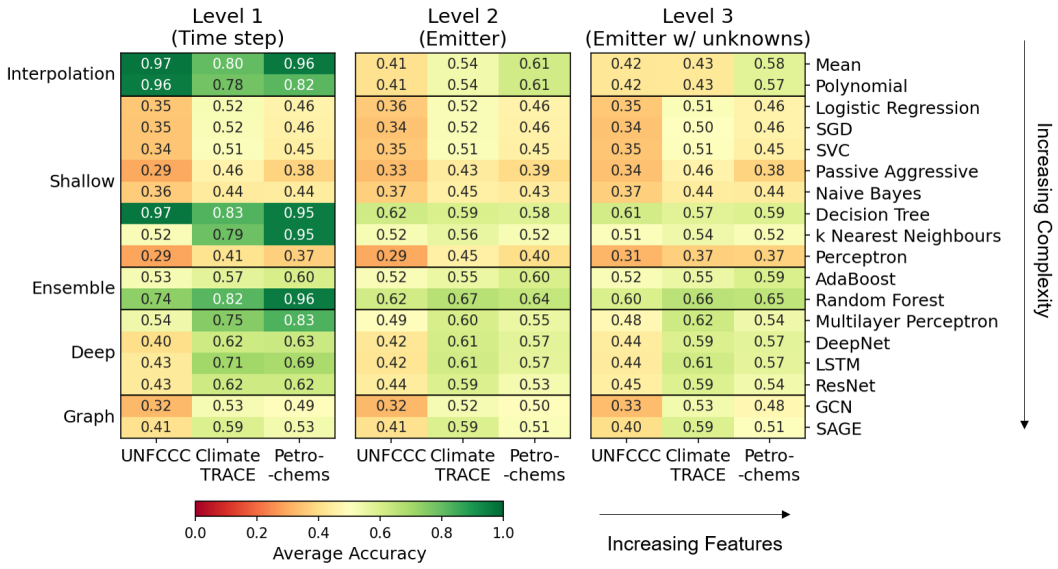

Figure 1: Average accuracy scores. Random guessing would result in 0.25 average accuracy. Trainable models were run for a maximum of 100 epochs. See SI2.1 for $F1$-score.

This study used uniform hyperparameter tuning for consistency, but performance for neural network based models could likely be increased with network structure adjustments and parameter tuning. In real-life a 70%/30% train/test split may be unrealistic. Further testing showed that average accuracies for the best performing models in each category reach within 10%, 20% and 20% of their peak accuracies with just 20% of training data available for levels 1, 2 and 3 respectively. A detailed figure is available in SI2.2. Figure 2 shows feature importance outputs. Figures 2a and 2b show that 'Category' for the UNFCCC and 'Capacity' for ClimateTRACE are the most valuable features for data collection. Figure 2c presents that if data for emitter 0 is not accessible, the best locations to gain data to improve knowledge of emitter 0's emissions are emitters 1, 2 and 4.

4 Discussion

Can ML provide an automated solution to GHG emission database completion? This study shows that ML classifiers can effectively and scalably complete emissions datasets in some cases. However, different models are appropriate for different types of gap-filling problem and complex models are not the best solution in many cases. Industrial ecologists and emissions analysts should not rush to use ML if their problem is not suitable. See SI3 or `https://hackmd.io/@luke-scot/ML-for-GHG-database-completion` for a guide for deciding which models should be used in which circumstances.

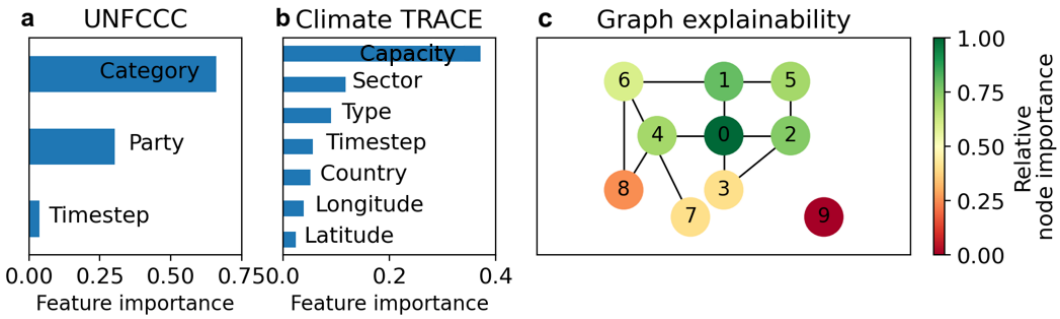

Figure 2: Model explainability representations. Relative feature importance for the decision tree classifier in the level 2 gap-filling problem are shown for: a) UNFCCC, b) ClimateTRACE. Plot c) shows the node importance analysis for node 0 using GNNExplainer (39) for the GCN method.

With only 2 features, ML models do not perform well on completing the UNFCCC dataset. If the UNFCCC database was associated with additional features, this may present an opportunity to enhance UNFCCC database coverage and target the tightening of UNFCCC regulation to those parties and features which are the most informative for the overall dataset. This will be the subject of future work.

Data collection at all emitters will continue to be the most accurate method but this is often not possible. Feature importance outputs may allow for a targeted approach to reduce the burden on organisations, such as ClimateTRACE, seeking carbon footprint transparency, as they could maintain data completeness without explicitly collecting all data points. The ability to infer reliable data across unobserved emitters may be a useful filter in detecting possible misreporting. If the predicted value is significantly different to the value reported, this could suggest some verification is required.

The ability to predict emissions categories without imposing any knowledge of the relationship between input features and output emissions permits easy incorporation of available data sources and avoids many of the biases involved in manual emissions estimates. In effect, any data source can be added to a learning-based model which could could open avenues for incorporation of many data types into a single framework for emissions estimates. Graph-based frameworks may be particularly helpful for incorporating material flow analysis and life-cycle assessment data due to their graphical forms. Further afield, a multimodal framework could be a basis for the merging of 'top-down' satellite measurements with 'bottom-up' emissions reports, as sought out in the remote sensing community (40).

Poor performance in scenarios with insufficient input features or high levels of complexity show the limitations of the techniques in this paper. Furthermore, results are based on 4-class classification which is too imprecise for many problems, and uncertainty is not quantified. To address these limitations, our future research will aim to quantify uncertainty across multimodal inputs and output an uncertainty distribution of values. Initial training time for all models in this study was less than 1 hour on a 16GB RAM computer. Standard neural networks require some level of re-training for each new data point which may become impractical with a regularly updated system. Graph frameworks are able to locally incorporate new data without re-training the whole network and may therefore be a more viable solution for incrementally improving models across large multimodal networks.

In conclusion, ML can be used to automate the completion of greenhouse gas emissions databases when enough input features are available and feature importance outputs can guide future data collection. Graph-based frameworks are particularly promising for future multimodal frameworks that could be used at the intersection of remote sensing and industrial ecology.

Acknowledgments and Disclosure of Funding

This work was supported by the UKRI Centre for Doctoral Training in Application of Artificial Intelligence to the study of Environmental Risks [grant number EP/S022961/1].

References

[1] ClimateTRACE *Manufacturing emissions data*, 2023. **URL**: `https://climatetrace.org/downloads`.

[2] UNFCCC, "Greenhouse gas inventory data - flexible queries annex I parties," 2023. **URL**: `https://di.unfccc.int/flex_annex1`.

[3] J. Rogelj, M. Den Elzen, N. Höhne, T. Fransen, H. Fekete, H. Winkler, R. Schaeffer, F. Sha, K. Riahi, and M. Meinshausen, "Paris agreement climate proposals need a boost to keep warming well below 2 c," *Nature*, vol. 534, no. 7609, pp. 631–639, 2016.

[4] T. Erb, B. Perciasepe, V. Radulovic, and M. Niland, "Corporate climate commitments: The trend towards net zero," in *Handbook of Climate Change Mitigation and Adaptation*, pp. 2985–3018, Springer, 2022.

[5] K. L. Christiansen, F. Hajdu, E. P. Mollaoglu, A. Andrews, W. Carton, and K. Fischer, ""our burgers eat carbon": Investigating the discourses of corporate net-zero commitments," *Environmental Science & Policy*, vol. 142, pp. 79–88, 2023.

[6] J. Arnold and P. Toledano, "Corporate net-zero pledges: The bad and the ugly," *Jack Arnold & Perrine Toledano, Corporate Net-Zero Pledges: The Bad and the Ugly,(2021) Columbia Center on Sustainable Investment Staff Publications*, 2021.

[7] IPCC, "Climate Change 2021: The Physical Science Basis. Contribution of Working Group I to the Sixth Assessment Report of the Intergovernmental Panel on Climate Change [Masson-Delmotte, V., P. Zhai, A. Pirani, S.L. Connors, C. Péan, S. Berger, N. Caud, Y. Chen, L. Goldfarb, M.I. Gomis, M. Huang, K. Leitzell, E. Lonnoy, J.B.R. Matthews, T.K. Maycock, T. Waterfield, O. Yelekçi, R. Yu, and B. Zhou (eds.)].," *Cambridge University Press.*, 2021.

[8] EPA, "Flight: Facility level information on greenhouse gases tool," *Environmental Protection Agency*, 2022.

[9] J. Marlowe and A. Clarke, "Carbon accounting: A systematic literature review and directions for future research," *Green Finance*, vol. 4, no. 1, pp. 71–87, 2022.

[10] DEFRA, "Guidance on how to measure and report your greenhouse gas emissions," *Department for Environment, Food and Rural Affairs*, 2009. **URL:** `https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/69282/pb13309-ghg-guidance-0909011.pdf`.

[11] EPA, "Rulemaking notices for ghg reporting," *Greenhouse Gas Reporting Program (GHGRP)*, 2022. **URL:** `https://www.epa.gov/ghgreporting/rulemaking-notices-ghg-reporting`.

[12] Ecoinvent, "Ecoinvent life cycle inventory database," 2022. **URL:** `https://ecoinvent.org/the-ecoinvent-database/`.

[13] E. B. de Souza Leao, L. F. M. do Nascimento, J. C. S. de Andrade, and J. A. P. de Oliveira, "Carbon accounting approaches and reporting gaps in urban emissions: An analysis of the greenhouse gas inventories and climate action plans in brazilian cities," *Journal of cleaner production*, vol. 245, p. 118930, 2020.

[14] L. Cullen, F. Meng, R. Lupton, and J. Cullen, "Reducing greenhouse gas emissions uncertainties for chemical production (under consideration)," 2023.

[15] F. Meng, L. Cullen, R. Lupton, and J. Cullen, "Greenhouse gas emissions from global petrochemical production (1978-2050) (under consideration)," 2023.

[16] T. Jusselme, E. Rey, and M. Andersen, "An integrative approach for embodied energy: Towards an lca-based data-driven design method," *Renewable and Sustainable Energy Reviews*, vol. 88, pp. 123–132, 2018.

[17] T. Potrč Obrecht, M. Röck, E. Hoxha, and A. Passer, "Bim and lca integration: A systematic literature review," *Sustainability*, vol. 12, no. 14, p. 5534, 2020.

[18] K. Stadler, R. Wood, T. Bulavskaya, C.-J. Södersten, M. Simas, S. Schmidt, A. Usubiaga, J. Acosta-Fernández, J. Kuenen, M. Bruckner, *et al.*, "Exiobase 3: Developing a time series of detailed environmentally extended multi-regional input-output tables," *Journal of Industrial Ecology*, vol. 22, no. 3, pp. 502–515, 2018.

[19] S. Pauliuk, N. Heeren, M. M. Hasan, and D. B. Müller, "A general data model for socioeconomic metabolism and its implementation in an industrial ecology data commons prototype," *Journal of Industrial Ecology*, vol. 23, no. 5, pp. 1016–1027, 2019.

[20] Climate TRACE, "Bringing radical transparency to global emissions," 2021. **URL:** `https://www.climatetrace.org/`.

[21] K. R. Gurney, J. Liang, R. Patarasuk, Y. Song, J. Huang, and G. Roest, "The vulcan version 3.0 high-resolution fossil fuel co2 emissions for the united states," *Journal of Geophysical Research: Atmospheres*, vol. 125, no. 19, p. e2020JD032974, 2020.

[22] Z. J. Steinmann, A. Venkatesh, M. Hauck, A. M. Schipper, R. Karuppiah, I. J. Laurenzi, and M. A. Huijbregts, "How to address data gaps in life cycle inventories: a case study on estimating co2 emissions from coal-fired electricity plants on a global scale," *Environmental science & technology*, vol. 48, no. 9, pp. 5282–5289, 2014.

[23] J. Pascual-González, C. Pozo, G. Guillén-Gosálbez, and L. Jiménez-Esteller, "Combined use of milp and multi-linear regression to simplify lca studies," *Computers & Chemical Engineering*, vol. 82, pp. 34–43, 2015.

[24] P. Hou, J. Cai, S. Qu, and M. Xu, "Estimating missing unit process data in life cycle assessment using a similarity-based approach," *Environmental science & technology*, vol. 52, no. 9, pp. 5259–5267, 2018.

[25] B. Zhao, C. Shuai, P. Hou, S. Qu, and M. Xu, "Estimation of unit process data for life cycle assessment using a decision tree-based approach," *Environmental science & technology*, vol. 55, no. 12, pp. 8439–8446, 2021.

[26] R. C. Lupton and J. M. Allwood, "Incremental material flow analysis with bayesian inference," *Journal of Industrial Ecology*, vol. 22, no. 6, pp. 1352–1364, 2018.

[27] J. Dong, J. Liao, X. Huan, and D. Cooper, "Expert elicitation and data noise learning for material flow analysis using bayesian inference," *Journal of Industrial Ecology*, 2023.

[28] H. Arbabi, M. Lanau, X. Li, G. Meyers, M. Dai, M. Mayfield, and D. Densley Tingley, "A scalable data collection, characterization, and accounting framework for urban material stocks," *Journal of Industrial Ecology*, vol. 26, no. 1, pp. 58–71, 2022.

[29] X. Vilaysouk, S. Saypadith, and S. Hashimoto, "Semisupervised machine learning classification framework for material intensity parameters of residential buildings," *Journal of Industrial Ecology*, vol. 26, no. 1, pp. 72–87, 2022.

[30] M. Algren, W. Fisher, and A. E. Landis, "Machine learning in life cycle assessment," in *Data Science Applied to Sustainability Analysis*, pp. 167–190, Elsevier, 2021.

[31] A. Ghoroghi, Y. Rezgui, I. Petri, and T. Beach, "Advances in application of machine learning to life cycle assessment: a literature review," *The International Journal of Life Cycle Assessment*, pp. 1–24, 2022.

[32] F. Donati, S. M. Dente, C. Li, X. Vilaysouk, A. Froemelt, R. Nishant, G. Liu, A. Tukker, and S. Hashimoto, "The future of artificial intelligence in the context of industrial ecology," *Journal of Industrial Ecology*, vol. 26, no. 4, pp. 1175–1181, 2022.

[33] R. He, L. Luo, A. Shamsuddin, and Q. Tang, "Corporate carbon accounting: a literature review of carbon accounting research from the kyoto protocol to the paris agreement," *Accounting & Finance*, vol. 62, no. 1, pp. 261–298, 2022.

[34] E. Süli and D. F. Mayers, *An introduction to numerical analysis*. Cambridge university press, 2003.

[35] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, *et al.*, "Scikit-learn: Machine learning in python," *the Journal of machine Learning research*, vol. 12, pp. 2825–2830, 2011.

[36] O. Sagi and L. Rokach, "Ensemble learning: A survey," *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, vol. 8, no. 4, p. e1249, 2018.

[37] I. Goodfellow, Y. Bengio, and A. Courville, *Deep learning*. MIT press, 2016.

[38] W. L. Hamilton, "Graph representation learning," *Synthesis Lectures on Artifical Intelligence and Machine Learning*, vol. 14, no. 3, pp. 1–159, 2020.

[39] Z. Ying, D. Bourgeois, J. You, M. Zitnik, and J. Leskovec, "Gnnexplainer: Generating explanations for graph neural networks," *Advances in neural information processing systems*, vol. 32, 2019.

[40] ESA, "Sentinel-1 SAR user guide," 2021.