Data Gap and Data Quality Analysis Final Report

Submitted by:

ERG & NREL

November 8, 2024

Contents

1. Background

In the support of the National Renewable Energy Laboratory (NREL), ERG is serving to improve public life cycle dataset repositories on the Federal LCA Commons (FLCAC). One of the tasks is to identify data gaps and data quality issues in the repositories on the FLCAC. The FLCAC contains publicly available data repositories that cover multiple industries including construction, energy, manufacturing, and transportation in the U.S. The FLCAC has been used by LCA practitioners as a reliable resource for free and publicly available Life Cycle Inventory (LCI) data.

This effort prioritizes data to support the Federal Buy Clean Initiative, with a particular emphasis on background data for materials of primary concern: steel, cement and concrete, asphalt, and flat glass. LCI background data are inventory data that are commonly found in the supply chains of products and services modeled in LCA studies. For these materials of primary concern, these background data include construction materials, energy, and their upstream inputs. The quality and coverage of these background datasets affect the accuracy of the final LCA results. To ensure the highest quality of the data on FLCAC, ERG is developing a research plan to identify potential data gaps and investigate possible data quality issues in these background data. The outcome of the data gap and data quality analysis will be used to inform priorities for data acquisition from FLCAC stakeholders.

In this study, a data gap refers to missing data or data that cannot sufficiently cover background data needs from the FLCAC. Therefore, a process that is covered by that FLCAC, but is of poor data quality, can be defined as a data gap. Therefore, the results from the data quality analysis can be used to inform the data gap assessment.

2. Methodology

This section details the methodologies used in this study. In section 2.1, we introduce the FLCAC data under study. Then, we describe the methods used for identifying hotspots in the supply chain of the materials covered by the Federal Buy Clean Initiative in section 2.2. The results from the hotspot analysis are used to identify data gaps in FLCAC and provide a priority list for data quality analysis. In Section 2.3, we introduce the indicators and metrics used in the data quality analysis. The results from the data quality analysis are also used as a guide to identifying data gaps; existing data with poor data quality information are flagged as data gaps.

2.1 Data Sources

In this study, we analyzed data from the FLCAC data repository. The below datasets were investigated:

- US Electricity Baseline (v.1.2020-08.0)
- NREL/USLCI (v.1.2024-10.0)
- MTU Asphalt Pavement Framework (v.1.2021-10.0)
- Heavy Equipment Operation (v.1.2022-04.0)
- CORRIM/Forestry and Forest Products (v.1.2019-12.0)
- NREL/Coal Extraction (v.1.2018-08.0)
- US Forest Service Forest Products Laboratory/Forestry and Forest Products (Woody Biomass) (v.1.2019-11.0)
- Construction and Demolition Debris (CDD) Management (v.1.2023-04.0)

The main data sources in this study include FLCAC data and environmentally extended input-output data. The FLCAC data include process-based life cycle inventory data from all available FLCAC repositories; each process's metadata, inputs and outputs, data quality information, and other relevant records were compiled programmatically. The input-output data are sourced from EPA's U.S. Environmentally Extended Input Output (USEEIO) model[1].

2.2 Hotspot Analysis for Data Gap Identification

We used structural path analysis (SPA) to identify the hotspot nodes for construction sectors in the IO model. A SPA model is based on input-output LCA models, and it is used to show the path relationships between different industry sectors. The advantage of using the SPA analysis is that it can separately list impacts by tiers (a.k.a. stages) and nodes (sector) on the supply chain of a production. The tiers or stages are the levels of inputs to the final production; a node (sector) on tier 1 is the input to the node on tier 0, which is the final production stage, and a node on tier 2 is the input to a node on tier 1. With different tiers and nodes separated, the details of inputs on the supply chain can be seen. Therefore, the results can be used to better identify hotspots in the background data.

In this study, we used the SPA modeling tool pyspa (https://github.com/hybridlca/pyspa), with the following parameters:

- Input-output model: USEEIO model v2.0.1, with 411 industries.
- Maximum level of tiers: four. This means the SPA model returns values from the first four tiers (or stages).
- Cut-off value: 0.1% of total impacts. This means that any values on each end node that are smaller than 0.1% of total impacts are not included.
- Impact category: GHG, Energy Use and Smog[2]. These three impact categories were analyzed individually.

Figure 1 provides an example of the result of using Energy Use as the impact category. Each box represents a node on the supply chain of the cement manufacturing industry. For example, *Natural Gas Distribution* is a node on Stage 1, it indicates that the *Natural Gas Distribution* sector is a direct input for *Cement Manufacturing* sector; *Oil and Gas extraction* sector, as an input to the *Natural Gas Distribution* sector, shows on Stage 2. No nodes on stage 4 contribute more than the cutoff value of total energy use, therefore none are shown in the result. Because the SPA results separately show the direct emissions on each node, a child node on a lower stage can have greater impact values than its parent node. In this study, we only consider the nodes that are above the cutoff threshold, which means the parents nodes with impacts below the cutoff threshold are not included as hotspots. In Figure 1, the parent nodes that do not meet the threshold are marked in orange.

We assessed seven industrial sectors that were most closely aligned with the priority Buy Clean Materials: *Cement manufacturing, Ready-mix concrete manufacturing, Glass and glass product manufacturing, Asphalt paving mixture and block manufacturing, Asphalt shingle and coating materials manufacturing, Iron and steel mills and ferroalloy manufacturing, Steel product manufacturing from*

[1] Ingwersen, W. W., Li, M., Young, B., Vendries, J., & Birney, C. (2022). USEEIO v2. 0, the US environmentally-extended input-output model v2. 0. *Scientific Data*, *9*(1), 194.
[2] Characterization factors from TRACI2.1 was used by the USEEIO model v2.0.1.

purchased steel. The supply chain hotspots of these sectors were summarized and ranked based on the stages (tiers) they appear in. This list showed areas within the supply chains of Buy Clean products that have significant environmental impact on GHG emissions, energy use and smog emissions. Processes classified under these industry sectors should be included in the background data repository for Buy Clean products. As a result, the list was used as a reference to identify data gaps in the FLCAC.

Figure 1: SPA analysis results for $1 of cement manufacturing, for total energy use; the orange boxes are parent nodes with direct energy use less than the cut-off threshold of 0.5%.

To identify data gaps in the FLCAC, the processes in the FLCAC were mapped to IO industries by each process' NAICS code provided in the metadata. Each hotspot industry identified in the SPA analysis which does not have one or more matching processes in the FLCAC was categorized as a potential data gap. For sectors that had mapped processes in the FLCAC data, we further inspected the processes to understand their potential similarities. For example, there were many processes that are categorized as organic chemicals in the FLCAC, which can be mapped to *Other Organic Chemical Manufacturing* sector. We

further inspected these processes to determine if any of them were relevant to the production of the construction materials, specifically. Process-based data collected for the construction processes were used to further check the presence of data gaps. There were industry sectors in the EEIO models that were too aggregated, such as *Other Basic Inorganic Chemical Manufacturing*. Therefore, additional work was required to understand the details. To further understand these hotspots and identify the data gaps, we used information from process-based LCA studies as a reference. We extracted data from 8 processes that best represent the Buy Clean materials from ecoinvent (v3.8) and calculated these processes' GHG, energy use, and smog impacts using TRACI (v3.8). We assessed and ranked the impacts from direct purchases, then compared these results to those identified in the SPA to better understand the specific makeup of the purchases made by these sectors. For example, ammonia was an input for the process *Portland Cement Production (US)* in ecoinvent. This input aligns with the SPA which identified the *Other Inorganic Chemical Manufacturing* sector as a potential hotspot, but provides much greater detail as to the specific chemical.

2.3 Data Quality Analysis

The data quality analysis in this study focuses on process-level data quality indicators. The data quality indicators are based on a data quality table (Table 1), a data quality score matrix (Table 2), and a LCIA compatibility assessment we developed partially based on EPA's Guidance on Data Quality Assessment for Life Cycle Inventory Data[3]. The data quality table (Table 1) provides 11 criteria, each criterion can be assessed from the process's metadata. The data quality pedigree matrix (Table 2) includes two criteria that can be assessed with a score; as shown in Table 2, a lower score (e.g. score 1) indicates better data quality than a worse or unknown score (e.g. score 5). Information on the Reviewed and Reproducibility criteria can be assessed based on expert judgement on the information gathered from the process's metadata. The Elementary Flow Completeness for LCIA criterion can be assessed by using the elementary flow list. The LCIA compatibility assessment contains three parts. The first two parts use metadata information in the FLCAC process to check if the processes are designed to cover certain levels of impact categories. This information includes the "completeness" statement and the "Intended Application" statement. Both statements use descriptive language to indicate if the dataset is designed to be used for certain impact categories. An example of such language is "...this unit process can be used for a full range of LCIA impact categories. The original study results were analyzed using the TRACI LCIA factors." The third part of the LCIA compatibility assessment is to analyze the number of elementary flows that can be mapped to certain impact categories. An example can be seen in Table 3.

We use a process in the USLCI database, "*Transportation, Train, Deisel Powered*", as an example to show how the table is completed. The results for the data quality table can be seen in the last column in Table 1. The scores for the pedigree matrix are 5 for Reviewed (No documented reviewed) and 3 for Reproducibility (the model was based on partially transparent data and based on publicly available data).

Note that in this study, a weighed final score to represent the overall data quality were not developed because as a publicly available dataset, the FLCAC might be used by various stakeholders in different applications. Different applications might require different weight factors to describe a total quantitative

[3] Edelen, A. AND W. Ingwersen. Guidance on Data Quality Assessment for Life Cycle Inventory Data. U.S. Environmental Protection Agency, Washington, DC, EPA/600/R-16/096, 2016.

data quality score. For example, if the application emphasizes using publicly available data, the indicator Reproducibility could be given a higher weighting factor.

Table 1: Criteria for process-level data quality description

Dataset attribute	Response option(s)	Example: Transport, train, diesel powered process in USLCI
Dataset date developed	Document (include start and end date if available)	12/31/2000
Dataset time period	Data is indicated to be representative of what time period	2000 – 2000
Unit process or system process	Document	Unit Process
Technology level of resolution	List identified process technology	Combustion of diesel in a locomotive
Geography level of resolution	List identified process geography	United States
FEDEFL aligned for interoperability	Yes/No, indicate # unmapped flows	Yes, 1 linked unit process, 8 elementary flows, 0 unmapped flows
Metadata documentation	Indicate which metadata fields in openLCA completed (General Information, Admin Info, Modeling and Validation)	No information on modeling and validation
Data quality scoring inclusion	Indicate whether included and % of flows available for	Not included
Uncertainty information inclusion	Indicate whether included and % of flows available for	Not included
Background LCI databases used as input (e.g., USLCI, GaBI)	Indicate dependencies on background databases	USLCI
Completeness of supply chain (background) data	How many cutoff flows	0

Table 2: Process-level data quality scores

Indicator	Definition	← Best Score				Lowest Score →
		1	**2**	**3**	**4**	**5 (default)**
Reviewed[4]	Assesses whether data have been independently quality assured and reviewed by subject matter and LCA experts.	Documented reviews by a minimum of two types of third-party reviewers	Documented reviews by a minimum of two types of reviewers, with one being a third-party	Documented review by a third-party reviewer	Documented review by an internal reviewer	No documented review
Reproducibility	Indicates transparency of the underlying model; Assesses the documentation of the source of the original data inputs to ; enable the independent recreation of the data by a third-party.	Underlying model and associated calculations are fully transparent and based on publicly available sources.	Underlying model and associated calculations are fully transparent and based on non-publicly available sources.	Underlying model and associated calculations are partially transparent and based on publicly available sources.	Underlying model and associated calculations are partially transparent and based on non-publicly available sources.	Underlying model and associated calculations are not transparent and based on non-publicly available sources or unknown.

[4] The criterion is based on the criterion provided in Guidance on Data Quality Assessment for Life Cycle Inventory Data: Edelen, A. AND W. Ingwersen. Guidance on Data Quality Assessment for Life Cycle Inventory Data. U.S. Environmental Protection Agency, Washington, DC, EPA/600/R-16/096, 2016.

Table 3: Elementary flow list for, Diesel Powered - RNA

Elementary Flow	Global Climate Air	HH Particulate Air	Acidification Air	Photochemical Oxidation Formation Air	Ozone Depletion Air
Carbon Dioxide	x				
Carbon Monoxide		x		x	
Methane	x			x	
Nitrogen dioxides		x	x	x	
Nitrous Oxide			x		
Particulate matter, ≤ 10µm					
Sulfur Oxides			x		
Volatile Organic Compounds					
Count (total number of flows)	**2**	**2**	**3**	**3**	**0**

3. Results

3.1 Data Gaps from Hotspot Analysis

For each of the three impact categories, seven sectors (*Cement manufacturing, Ready-mix concrete manufacturing, Glass and glass product manufacturing, Asphalt paving mixture and block manufacturing, Asphalt shingle and coating materials manufacturing, Iron and steel mills and ferroalloy manufacturing, Steel product manufacturing from purchased steel*) were analyzed in the SPA model separately, the results are aggregated by the stage and sector name. We found 49 sectors as hotspots from the three impact categories. Table 4 shows the results from the GHG impact category. Results show that most of the sectors identified cover all three stages, but a few sectors only show as nodes at one stage, such as Air Transportation (only stage 1) and Grain Farming (only stage 2). These sectors can be used differently when identifying data gaps. For example, Grain Farming only shows as a stage 2 node in two materials, it indicates that the sector is not a direct emission or energy use contributor, thus it can be given lower-level priority when identifying data gaps.

Table 4: SPA analysis results for the selected seven industries for GHG emissions

Sector Name	IO code	Stage
Air transportation	481000	Stage 1
Alumina refining and primary aluminum production	331313	Stage 1, 2
Asphalt paving mixture and block manufacturing	324121	Stage 1
Asphalt shingle and coating materials manufacturing	324122	Stage 1
Cement manufacturing	327310	Stage 1,2,3
Clay product and refractory manufacturing	327100	Stage 1

Sector Name	IO code	Stage
Coal mining	212100	Stage 1,2,3
Copper, nickel, lead, and zinc mining	212230	Stage 1,2,3
Electric power generation, transmission, and distribution	221100	Stage 1,2,3
Glass and glass product manufacturing	327200	Stage 1,2
Grain farming	1111B0	Stage 2
Ground or treated mineral and earth manufacturing	327992	Stage 1
Industrial gas manufacturing	325120	Stage 1
Iron and steel mills and ferroalloy manufacturing	331110	Stage 1,2,3
Iron, gold, silver, and other metal ore mining	2122A0	Stage 1,2,3
Lime and gypsum product manufacturing	327400	Stage 1,2,3
Mineral wool manufacturing	327993	Stage 1
Miscellaneous nonmetallic mineral products	327999	Stage 1
Natural gas distribution	221200	Stage 1,2
Nonferrous metal (except aluminum) smelting and refining	331410	Stage 1,2
Oil and gas extraction	211000	Stage 1, 2, 3
Other basic inorganic chemical manufacturing	325180	Stage 1
Other basic organic chemical manufacturing	325190	Stage 1,2
Other durable goods merchant wholesalers	423A00	Stage 1
Other nonmetallic mineral mining and quarrying	2123A0	Stage 1,2
Other petroleum and coal products manufacturing	324190	Stage 1,2
Other support activities for mining	21311A	Stage 3
Paper mills	322120	Stage 1,2
Paperboard container manufacturing	322210	Stage 1
Paperboard mills	322130	Stage 2
Petrochemical manufacturing	325110	Stage 2
Petroleum refineries	324110	Stage 1
Pipeline transportation	486000	Stage 1,2,3
Plastics material and resin manufacturing	325211	Stage 1,2
Plastics packaging materials and unlaminated film and sheet manufacturing	326110	Stage 1
Rail transportation	482000	Stage 1,2
Scrap	S00401	Stage 1,2
Semiconductor and related device manufacturing	334413	Stage 1
Solid Waste Landfill	562212	Stage 1
Steel product manufacturing from purchased steel	331200	Stage 1
Stone mining and quarrying	212310	Stage 1
Truck transportation	484000	Stage 1,2,3
Water transportation	483000	Stage 1

The hotspot industries then were mapped to FLCAC using the method described previously. Table 5 shows some examples of the mapping between FLCAC processes and industry hotspots from the SPA analysis (the full list can be seen in the Data Gap Results Excel file).

Table 5: Excerpts from the mapping result

IO Sector	NAICS	Number of processes in FLCAC	Process name in FLCAC
Oil and gas extraction	2111: Oil and Gas Extraction	18	Crude oil, off-shore domestic, at extraction
			Crude oil, off-shore import, at extraction
			Crude oil, on-shore domestic, at extraction
			Crude oil, on-shore import, at extraction
			Crude oil, production mixture, at extraction
			Natural gas, at processing, conventional, kg
			Natural gas, at processing, shale, kg
Copper, nickel, lead, and zinc mining	2122: Metal Ore Mining	1	Bauxite, at mine
Glass and glass product manufacturing	3272: Glass and Glass Product Manufacturing	2	Glass fiber
			Glass fiber, reinforced polymer
Natural gas distribution	2212: Natural Gas Distribution	14	Natural gas extraction and processing – Uinta (and other 13 locations); Transportation, pipeline, natural gas
Rail transportation	4821: Rail Transportation	1	Transport, train, diesel powered
Iron and steel mills and ferroalloy manufacturing	3311: Iron and Steel Mills and Ferroalloy Manufacturing	13	Iron and steel, production mix
			Steel, billets, at plant
			Steel; cold rolled coil, at plant
			Steel, stainless 304, flat rolled coil
			Steel, stainless 304, scrap
			Steel; hot rolled coil, at plant
			Steel, stainless 304, quarto plate
			Steel; pickled hot rolled coil, at plant
			Steel; sections, at plant
			Steel, liquid, at plant
			Steel; hot-dip galvanised coil, at plant
			Steel; plate, at plant
			Metallurgical coke, at plant
Petrochemical manufacturing; Industrial gas manufacturing; Other basic inorganic chemical manufacturing; Other basic organic chemical manufacturing	3251: Basic Chemical Manufacturing	47	Ethanol, 85%, at blending terminal, 2022
			Ethylene glycol, materials production, organic compound, at plant, kg
			Hydrogen, liquid, synthesis gas, at plant
			Soy biodiesel, production, at plant
Clay product and refractory manufacturing	Process not available in FLCAC	Process not available in FLCAC	Process not available in FLCAC

The results show that for the selected construction materials, the coverage of background data is better in some industries, such as *Natural gas distribution*. In some other industries, there is only 1 process available (*Rail transportation*) or no process is available (*Clay product and refractory manufacturing*). In some cases, there are multiple processes available for a given industrial category, but these processes may not be relevant to the supply chains of the priority materials (e.g., Bauxite, at mine, is a process within the *Metal Ore Mining* category). Chemicals are the most complicated processes to analyze due to the highly aggregated nature of the input-output model and the metadata provided by the FLCAC processes. The four-digit NAICS code is the highest resolution industry code provided by each FLCAC process' metadata. Therefore, four industrial sectors (*Petrochemical manufacturing; Industrial gas manufacturing; Other basic inorganic chemical manufacturing; and Other basic organic chemical manufacturing*) were all mapped to the four-digit NAICS code 3251 (*Basic Chemical Manufacturing*), while 47 FLCAC processes are included in that category. Some of the chemical processes might not be construction materials or their upstream inputs; while some chemicals that are used in the construction industry are not covered, such as explosives. The results show that some industries are not covered by FLCAC, and some are covered but lack data in construction materials and their upstream inputs.

We then use ecoinvent data as a reference to identify the data gaps that might be missed from the SPA study due to the aggregation of the industries in the input-output model. With the results from both the input-output model and process-based model, we identified 18 data gaps in FLCAC. The results are summarized in Table 6. Some of these processes are hotspots from both the input-output model and the process-based model, such as *Rail transportation*; some processes are hotspots from only one LCA model, such as *Semiconductor and related device manufacturing* (EEIO) and *Heat from sweet gas* (ecoinvent). In some cases, the process in ecoinvent can explain the inclusion of an aggregated industrial sector. For example, Magnesium oxide is an input for cement production in ecoinvent, it could be the hotspot *Other basic inorganic chemical manufacturing* in the IO model. Note that some hotspots from the IO model are not flagged as data gaps because they are service-related sectors. These sectors were identified in the SPA analysis because of the economic exchange hotspots, the impacts associated were likely not from production. These sectors are: *Other durable goods merchant wholesalers, Scenic and sightseeing transportation and support activities for transportation,* and *Services to buildings and dwellings*.

Results show that some of the data gaps can come from the same product system, such as the *Iron and other ferro manufacturing* process and the *Iron ore mining* process. These processes should be given priority when fixing the data gaps. Some processes related to waste management are not included in the list in the data gap list, but they are hotspots in both the IO model and the process-based model.

Table 6: data gaps identified based on USEEIO model and ecoinvent processes.

Process	In FLCAC	Reference	IO Industry	Data Gap Notes
Clinker production	No	Ecoinvent/EEIO	Clay product and refractory manufacturing	
Cobalt	No	Ecoinvent/EEIO	Nonferrous metal (except aluminum) smelting and refining	
Cobalt mining	No	Ecoinvent/EEIO	Copper, nickel, lead, and zinc mining	Will be added to USLCI soon
Ferrochromium production	No	Ecoinvent/EEIO	Iron and steel mills and ferroalloy manufacturing	
Ferromanganese production	No	Ecoinvent/EEIO	Iron and steel mills and ferroalloy manufacturing	
Ferronickel production	No	Ecoinvent/EEIO	Iron and steel mills and ferroalloy manufacturing	
Flat glass	No	Ecoinvent/EEIO	Glass and glass product manufacturing	
Iron ore mining	No	Ecoinvent/EEIO	Iron, gold, silver, and other metal ore mining	
Iron pellet	No	Ecoinvent/EEIO	Iron and steel mills and ferroalloy manufacturing	
Magnesium oxide	No	Ecoinvent/EEIO	Other basic inorganic chemical manufacturing	
Mineral wool	No	EEIO	Mineral wool manufacturing	
Miscellaneous nonmetallic mineral	No	EEIO	Miscellaneous nonmetallic mineral products	
Pig iron	No	Ecoinvent	Iron and steel mills and ferroalloy manufacturing	
Pitch	No	Ecoinvent	Asphalt shingle and coating materials manufacturing	
Semiconductor	No	EEIO	Semiconductor and related device manufacturing	
Smelting and refining of nickel concentrate	No	Ecoinvent/EEIO	Nonferrous metal (except aluminum) smelting and refining	
Synthetic fuel production	No	Ecoinvent	Petrochemical manufacturing	
Zinc mining	No	Ecoinvent/EEIO	Copper, nickel, lead, and zinc mining	

3.2 Data Quality

Process Metadata

Based on the data quality criteria provided in section 2, we analyzed 625 identical processes that were identified as supply chain inputs for the Buy Clean materials based on the SPA analysis. The full list is shown in the Process Metadata Excel file attached to the report.

The results from the data quality analysis showed that most processes need improvements in their metadata. The metadata fields for FLCAC processes are often left blank, this includes data quality scoring inclusion (50% of processes have no data quality score), and uncertainty information inclusion (none of the processes have uncertainty information) fields. Some of these fields were filled with insufficient information. For example, the Data Collection Description field for the process Cardboard; AWC is "2012", which might indicate that the data was collected in 2012, but more description should be provided to reduce ambiguity.

Some processes are mapped to the wrong NAICS code, such as *Sulfur, at plant*, the process is mapped to 2111: *Oil and Gas Extraction*. There are some other examples in FLCAC, such as *Bituminous coal, combusted in industrial boiler, at pulp and paper mill (EXCL.)* mapped to *3221: Pulp, Paper, and Paperboard Mills*, the wrong mapping is usually because the process is a co-product or by-product and mapped to the NACIS code for the main product. This could be a data quality issue that we will discuss in the next section.

Cutoff Flows

The second data quality issue identified was the number of cutoff flows. Cutoff flows in FLCAC are technosphere flows that are used as inputs to produce other processes, but the cutoff flows' inventories are not available. Because they are known inputs for other processes, these cutoff flows' names are listed in the datasets as placeholders. The cutoff flows present in the activity data of a process indicate possible underestimation of the impacts and a reduced interoperability of the process. Among the 625 processes under study, there are 152 processes that include at least one cutoff flows as inputs (cutoff flows categorized as waste flows were not included). Figure 2 shows the distribution of cutoff flows in each process among the 152 processes that have non-waste flow cutoff flows. A full list of the cutoff flows that are in the inventories of the 625 processes can be found in the Process Metadata Excel file. Some major cutoff flows include:

- Scraps, recycled materials, and waste materials used as inputs. These cutoff flows include metal scraps such as steel scraps, aluminum scraps, and zinc scraps; recycled plastics including PVC; waste oil, waste tires, and waste used in energy recovery. The use of scraps in the inventory makes it more challenging for a standard allocation method to properly allocate the impacts from using these materials as inputs.
- Chemicals. These cutoff flows include chemicals that are used in construction materials such as titanium dioxide, sulfur dioxide, sodium sulfate, sodium hydrosulfite, sodium chlorate, pentane, monoethanolamine, latex, hydrogen peroxide. Some processes (Aluminum, cold rolling, at plant, Ethanol, denatured, switchgrass, biochemical) list a large amount of "Chemicals (unspecified)" or

"Chemicals organic, at plant" as one of the cutoff inputs; it introduces large uncertainty to the final impacts and should be addressed.

- Mineral products: These products are often used directly in the construction industry, some of them are identified as data gaps from the hotspot analysis. These products include clay, and zeolite.
- Energy: Some cutoff flows are energy from electricity, steam, or the combustion of fossil fuels. Some of these processes have large contributions and should be fixed to reduce uncertainty. These processes include Electricity from hydro, Propane combusted in equipment, and Steam, at plant. Note that steam could also be a by-product and could be treated as one of the recycled or waste flows used as inputs.
- Disposal processes: 61 cutoff flows are different kinds of disposal processes, including the disposal of chemical waste, coal ash, mineral and mining waste. Note that in processes' metadata, these processes were categorized as product flows rather than wastes flows, which were excluded from the study.

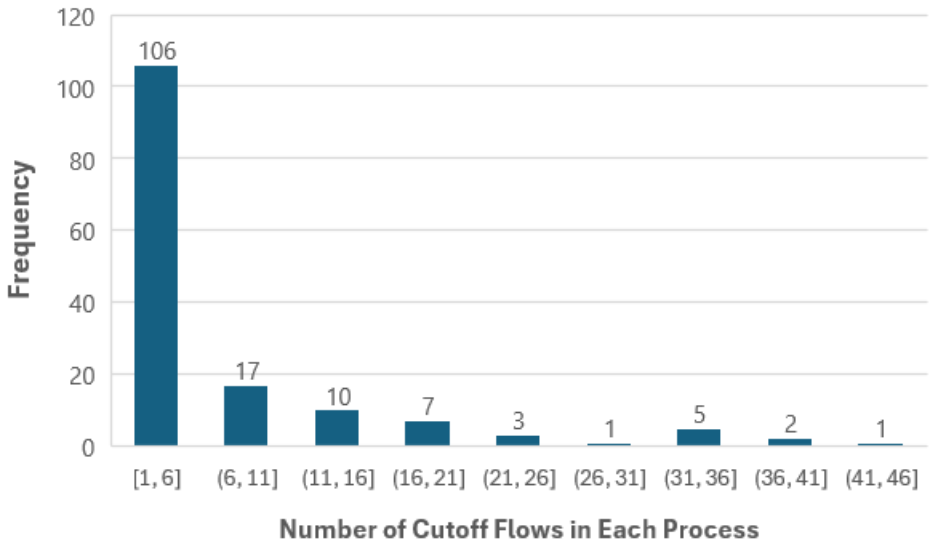

Figure 2: Distribution of number of cutoff flows in each process (total number of processes: 152)

Process-Level Data Quality Matrix Scores

Figure 3 shows the data quality scores for the 625 processes under study. Results show that for the indicator Reviewed, the 99% of the processes either get the best score (score 1) or the worst score (score 5). This is partially due to the poor documentation of the metadata. For the indicator Reproducibility, 65% of the processes were scored 5, indicating most of the processes' data are neither publicly available nor transparent.

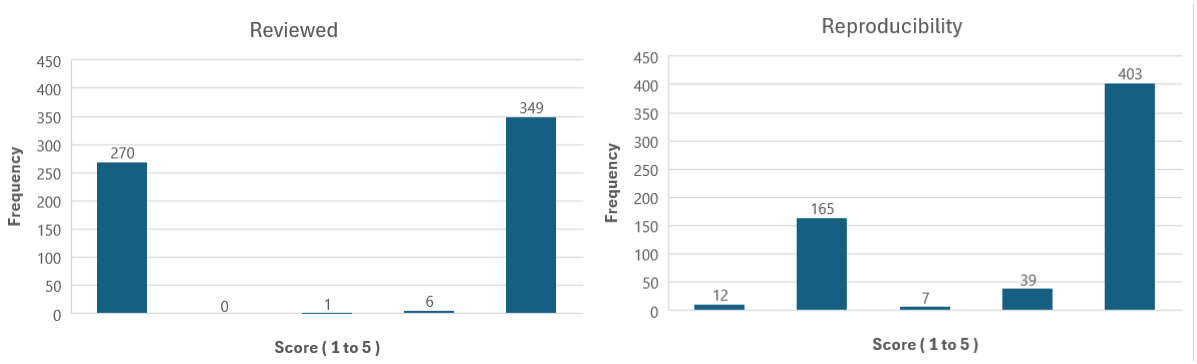

Figure 3: Distribution of reviewed and reproducibility scores (total number of processes: 625)

LCIA Coverage

Table *7* shows a summary of the intended application description to the LCIA methods intended to cover by the process. Results show that around 40% of the processes did not provide any information in the intended applications, while around 10% of the processes provided certain information but were irrelevant to the LCIA information. The other half of the processes provide indented application as "a full range of LCIA impact categories" and/or "use TRACI as the LCIA factors". Further data analysis on the number of TRACI substances covered showed that around one third (189) of the processes that have less than 10 substances covered, and 270 processes have no elementary flows covered by TRACI. These processes are mostly electricity production mix or consumption mix processes that are intended to have zero elementary flows in their inventories (not included in Table 8). These processes' impacts would be accounted for in their upstream inputs. However, a significant number of processes are not electricity production or consumption mix and have elementary flows that are not covered by any TRACI impact category. The list of these processes can be found in Table 8. Not including elementary flows and/or not covered by TRACI impact categories result in underestimating the impacts because no direct impacts could be captured by the impact assessment model. We include these processes as the final data gap list to inform the data curator to prioritize updating these processes.

Table 7: The Intended Application Information Provided by Process Metadata

Intended Application Description	Number of Processes
"To cover a full range of LCIA impact categories"	270
"To cover a full range of LCIA impact categories", "using the TRACI LCIA factors"	34
"TRACI 2.1 was the impact assessment method originally used for this study."	5
"Developed and analyzed with the ReCiPe impact assessment methods"	1
A description is provided but not related to LCIA methods intended to cover	59
Field left blank	256

Table 8: Processes with zero elementary flows captured by TRACI LCIA methods

Process Name	Number of Elementary Flows	Number of Product Flows	Number of Cutoff Flows
Processing of gypsum; milled; at drywall processing facility	6	25	
Processing of asphalt shingles; ground; at processing plant	6	19	
Potassium sulphate alum, as K2O, hardboard	5	7	
Calcium carbonate, ground, 20 micron, at plant	5	27	8
Calcium carbonate, ground, fine treated, 3 micron, at plant	5	39	12
Calcium carbonate, ground, 30 micron, at plant	4	24	6
Calcium carbonate, ground, fine slurry, 3 micron, at plant	4	39	12
Calcium carbonate, ground, screened grade, at plant	4	23	6
Slack wax; AWC	3	1	
Limestone, at mine	2	8	
Soy biodiesel, production, at plant	2	14	3
Thermoforming, rigid polypropylene part, at plant	2	16	2
Sodium chloride, at plant	1	7	
Hydrogen; proton exchange membrane water electrolysis; at plant	1	5	1
Hydrogen; solid oxide electrolysis, at plant	1	6	1
Oxygen, liquid, at plant	1	3	
Secondary bonding application, rigid composites part, at plant	1	9	2
Crude palm kernel oil, at plant	1	11	1
Natural gas, processed, for energy use, at plant		2	
Natural gas, processed, for material use, at plant		2	
Ethanol, 85%, blended, at service station, 2022		7	3
Lubricant feedstock, at refinery		2	
Petroleum refined, for energy use, at plant		2	
Petroleum refined, for material use, at plant		2	
Acrylonitrile; at plant		7	
Ethanol, 85%, at blending terminal, 2022		7	1
Ethanol, denatured, at refueling station, 2022		6	
Ethanol, denatured, mixed feedstocks, at conversion facility, 2022		6	
Wood chip pyrolysis; Tucker Renewable Natural Gas (RNG) thermochemical conversion process; at plant		6	
Carbon fiber, reinforced polymer; at plant		6	
Carbon fiber; at plant		4	
Epoxy; at plant		5	
Methyl methacrylate, MMA; at plant		5	
Polyacrylonitrile, PAN: at plant		7	
Packaging materials; SE		5	

Packaging wrapping material; at plant		4	
Glass fiber; at plant		6	2
Glass fiber, reinforced polymer; at plant		5	1
Aluminum ingot, production mix, at plant		3	
Aluminum, sheet, coated, at plant		5	1
Transport, ocean freighter, average fuel mix		3	
Transport, barge, average fuel mix		3	
Switchgrass, at conversion plant, 2022		5	
Corn stover, at conversion plant, 2022		5	
Corn grain, at conversion plant, 2022		5	
Forest residue, preprocessed, at conversion facility		3	
Wheat straw, at conversion plant, 2022		6	
Transport, pipeline, unspecified petroleum products		3	
Packaging and information sheets, i2900 desktop scanner		6	2
Scanner, packaging and information sheets		8	1
Electricity, at cogen, for natural gas turbine		3	
Electricity, bauxite mining regions		8	2
Electricity, nuclear, at power plant		2	
Electricity, aluminum smelting and ingot casting regions		7	1
Electricity, alumina refining regions		8	2

Outdated data

In this study, we found that a number of processes were generated a few decades ago but have not been updated. We used the "valid until" field in the metadata to filter such processes and used data quality indicators provided by EPA's Guidance on Data Quality Assessment as a reference to determine a threshold to flag old data as data gaps. In EPA's Guidance, flows that are greater than 15 years from the target year receive a score of 5. We used 15 years as the cutoff threshold and identified 228 processes that exceed this threshold. These processes include main construction materials such as limestone, Portland cement, steel (liquid and billets), and aluminum (primary or secondary); fuel or energy such as electricity and energy from the combustion of fossil fuels; transportation processes such as truck transportation processes; and chemicals such as ethanol.

3.3 Data Gaps List

The final data gap list consists of data gaps identified from the hotspot analysis, cutoff flows, processes with limited information on LCIA category coverages, and processes with poor temporal representativeness. The whole list of data gaps can be found in the Data Gap Results Excel file.

3.4 Comparison with EPA Data Gap Assessment

The U.S EPA conducted a data gap assessment to support its Label Program for Low Embodied Carbon Construction Materials[5]. This project identified data gaps provided by reviewing the data gaps provided in existing published PCRs and inputs from PCR committees, program operators, and relevant PCR committee members. The Applied Economics Office at National Institute of Standards & Technology (NIST)[6] also conducted a study to identify public data gaps for LCA modeling. Table *9* provides a crosswalk between EPA's results and data gaps identified in this project. Note that in the data gaps provided in Table 6 include the main construction materials as products, such as flat glass and concrete. These products are out of scope of EPA's data gap project, thus excluded from their data gap list. Results showed that there are only several processes or materials identified by both agencies. The main reasons are: 1) the method we used focused on hotspot analysis, which considers the hotspots on the supply chain based on certain impact categories (GHG, energy and smog in this study). Some data gaps identified by LCA practitioners and stakeholders, such as window hardware, were not captured as a hotspot for these three categories, but they might be hotpots when other categories are considered; 2) on the other hand, the hotspot analysis can capture more processes on the supply chain, including the second and more upstream tiers. The data gaps identified by the EPA project focus more on the first-tier inputs; 3) some data gaps we identified were aggregated commodities, in some cases no specific product can be identified, such as aggregates identified by EPA, in the hotspot analysis, these aggregates might be included in the Other basic inorganic chemical manufacturing or Other basic organic chemical manufacturing sectors; 4) in this project, we used data quality results to help identify data gaps, processes listed as cutoff flows, processes with no link to LCIA methods were flagged as data gaps and these data gaps might not be captured in the EPA's study; 5) there were some recent updates on the transportation data set in FLCAC, such freight barge and train transportation as these processes were captured as data gaps possibly after the EPA's study.

Table 9: Crosswalk of Data Gaps Identified

	Process/Material	**Data Gap Identified by EPA**	**Data Gap Identified by NIST**	**Data Gap Identified by this Project**
Raw Materials and Extraction	Aggregates	Yes	Yes	
	Iron ore	Yes	Yes	Yes
	Iron pellet			Yes
	Pig iron			Yes
	Steel product manufacturing from purchased steel			Yes
	Mn wear parts	Yes	Yes	Yes (as Miscellaneous nonmetallic mineral)

[5] US EPA, Label Program for Low Embodied Carbon Construction Materials https://www.epa.gov/greenerproducts/label-program-low-embodied-carbon-construction-materials

[6] NIST, Metrics and Tools for Sustainable Buildings https://www.nist.gov/programs-projects/metrics-and-tools-sustainable-buildings

	Hydrated lime	Yes		
	Steel slag	Yes	Yes	
	Clinker			Yes
	Clay		Yes	Yes (cutoff process)
	Cobalt and Cobalt mining			Yes
	Ferrochromium production			Yes
	Ferromanganese production			Yes
	Ferronickel production			Yes
	Mineral wool			Yes
	Hard coal mine operation			Yes
	Pitch			Yes
	Zinc mining			Yes
	Smelting and refining of nickel concentrate			Yes
	Tires	Yes	Yes	
Chemicals and Additives	Chemicals	Yes		Yes
	Magnesium oxide			Yes
	Concrete admixtures	Yes	Yes	
	Explosives	Yes	Yes	Yes
	Flocculants	Yes	Yes	
Transportation	Asphalt tankers (international)	Yes		
	Freight, barge	Yes		
	Freight, train	Yes		
	Overloaded asphalt trucks	Yes		
Fuels (consumption mixes)	Biofuels	Yes	Yes	
	CNG	Yes		
	Electric vehicles	Yes		
	Hydrogen	Yes		
	LNG	Yes		

	Propane	Yes		
	Natural gas distribution			Yes
	Synthetic fuel production			Yes
	Diesel			Yes (the process itself is not available but the LCI is provided in petroleum refinery processes.)
	Heavy fuel oil			Yes
	Renewable diesel	Yes	Yes	
Building materials	Glazing/novel coatings	Yes		
	Vinyl window framing	Yes		
	Window hardware	Yes		
Energy and resources	Crude oil extraction	Yes		
	Electricity mixes	Yes		
	REC methods	Yes		
Packaging	Cardboards	Yes		
	Plastics (more options)	Yes		
Other	Semiconductor	No		Yes
	Scraps (including metal scraps, wastes)	No		Yes
	Water	Yes	Yes	Yes
	Waste	Yes	Yes	Yes
	Grinding aids		Yes	?
	Hydraulic fluid		Yes	?
	Calcium chloride		Yes	Yes
	Hydrated lime		Yes	?
	Clay, Shale, kaolin, marl		Yes	Yes
	Gypsum, Sand, silica sand		Yes	Yes
	Refractory		Yes	Yes

4. Future Work

4.1 Stakeholder Engagement for Filling Data Gaps

In this project, we identified 18 data gaps from the hotspot analysis and 590 processes potential data gaps due to different data quality issues. Future work from the results of the study can focus on prioritizing filling the data gaps and curate data. One way for the data curation is to identify stakeholders, institutions, private companies and other agencies that might have such data. Table 10 provides some examples.

Table 10: Examples of Potential Agencies for Filling Data Gaps

Data Gap	Agency	Agency Type	Agency Website	Notes
Aluminum production	The Aluminum Association	US Industry Association	https://www.aluminum.org/	Potential industrial average LCI data
Steel, Iron	American Iron and Steel Institute	US Industry Association	https://www.steel.org/	Potential industrial average LCI data
Chemicals	American Chemistry Council	US Industry Association	https://www.americanchemistry.com/	Potential industrial average LCI data; possible activity data for construction materials
Mining activities	National Mining Association	US national trade organization	https://nma.org/	Potential industrial average LCI data

4.2 Disaggregation of Chemical Industries

We have identified that chemicals were used in supply chain of the construction materials; the results from EPA's project also confirmed the finding. However, neither the hotspot analysis nor the results from the stakeholder engagement from EPA's study provide many specific chemicals used in the construction industry. The disaggregation of chemical industries is crucial to better understand the data gaps and prioritize processes for data curation purposes. The future work of the project should include identifying or designing a method to understand the specific chemicals as hotspots in the construction industry.