Principal Component Analysis and Clustering for Scope 3 Greenhouse Gas (GHG) Emissions Estimation with Hierarchical Linear Modeling

Amber Jensen | August 21, 2025

Overview

- Introduction & Background
- Problem Statement & Research Question
- Data Preprocessing
- Principal Component Analysis (PCA)
- Hierarchical Clustering
- Hierarchical Linear Modeling (HLM) Results
- Conclusion

Introduction & Background

- Accurately estimating Scope 3 greenhouse gas (GHG) emissions is challenging due to inconsistent reporting standards, incomplete data, complex supply chains across diverse industries.

- This project developed five Hierarchical Linear Models (HLM) for each of 15 Scope 3 emissions datasets, using company financial and self-reported emissions data.

- HLM models provided strong predictive accuracy, and matched or outperformed published machine learning models, including Random Forest and AdaBoost models.

Problem Statement & Research Question

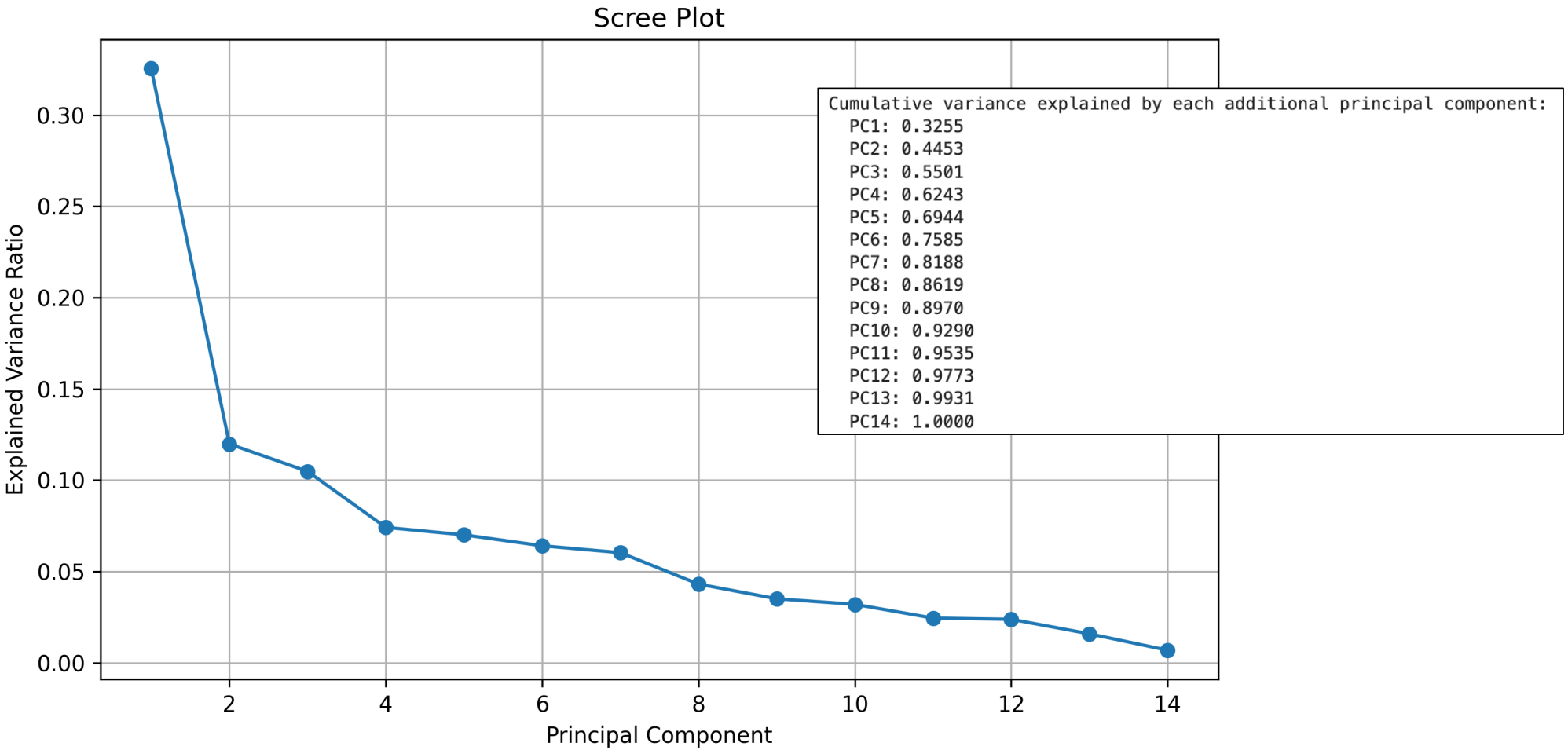
- **Problem Statement**: Scope 3 greenhouse gas (GHG) emissions are critical for understanding an organization's total emissions footprint. However, accurately estimating Scope 3 emissions is difficult because the data is highly complex and often incomplete due to challenges in collecting and standardizing data across complex supply chains.

- **Research Question**: Does integrating Principal Component Analysis (PCA) and hierarchical clustering with Hierarchical Linear Modeling (HLM) improve the accuracy of Scope 3 emissions predictions, while preserving the transparency and interpretability that HLM models offer?

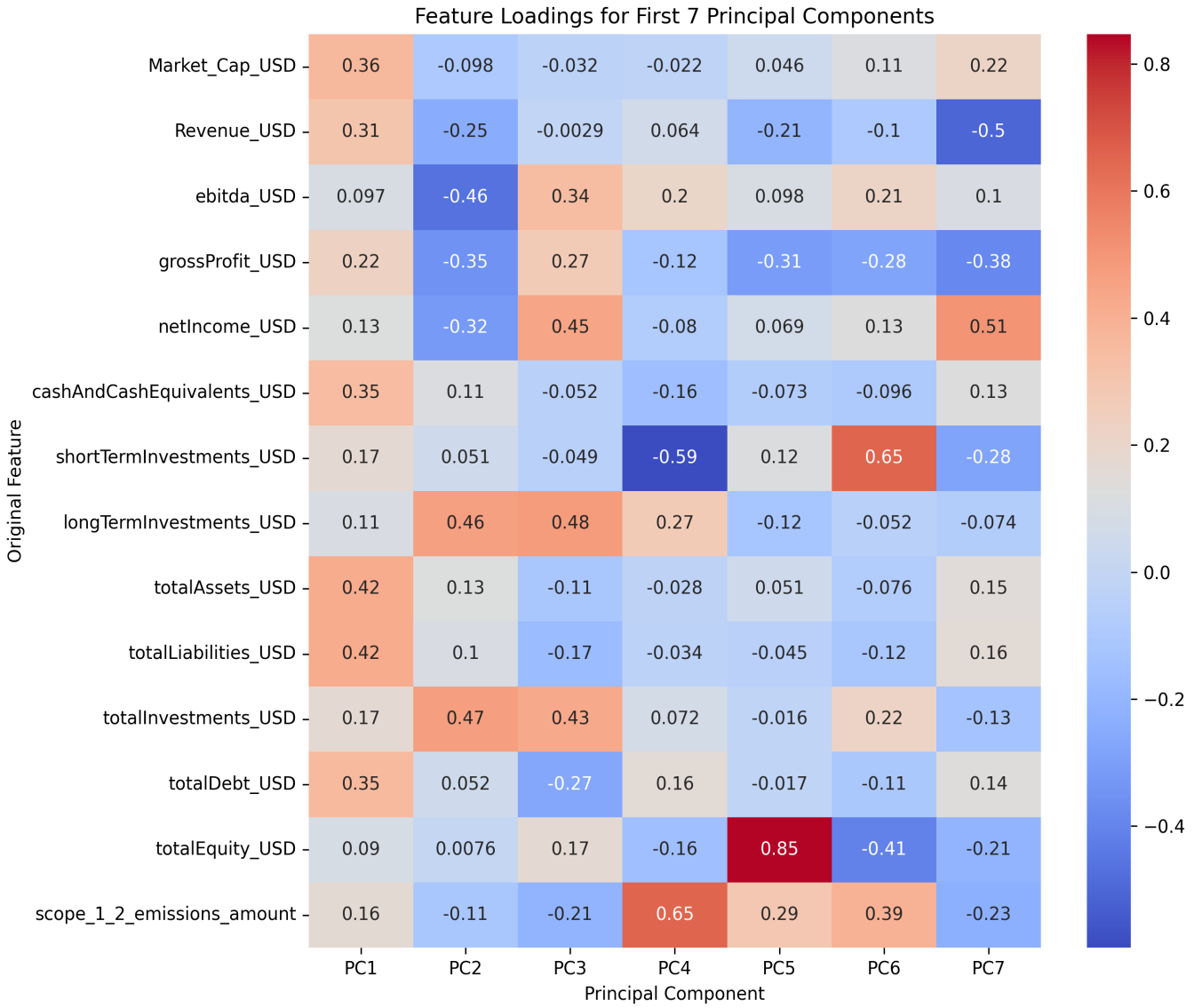
Data Preprocessing

- **Numeric Feature Selection:** Used relevant numeric columns only.
- **Signed Log Transformation:** Reduced skew; handled unit differences, negatives, and zeros.
- **Standardization:** Scaled features (z-score) before PCA.
- **Missing Data:** Removed incomplete rows.
- **Train/Test Split:** Trained on 2018–2022; tested on 2023 for known companies.
- **Scope 3 Datasets:** Analyzed 15 source datasets; Fuel & Energy shown as example.

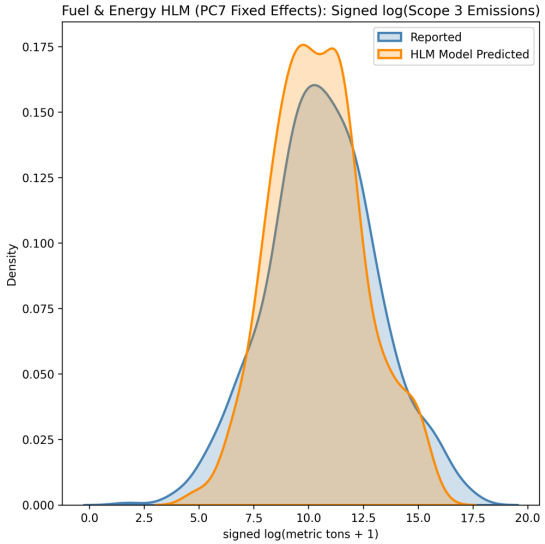
Scree Plot & Cumulative Explained Variance



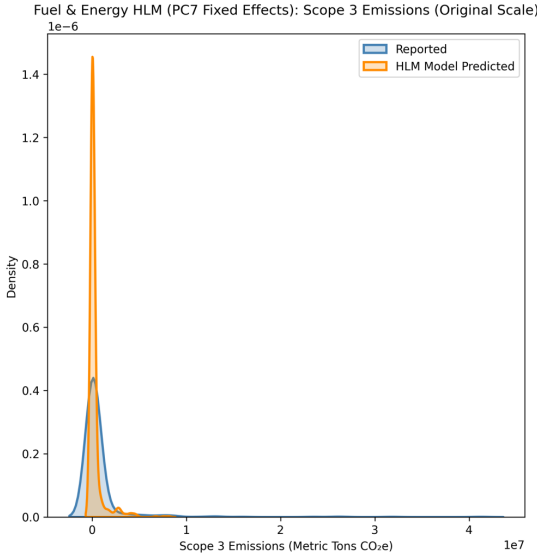
Feature Loadings Heatmap for First 7 Principal Components

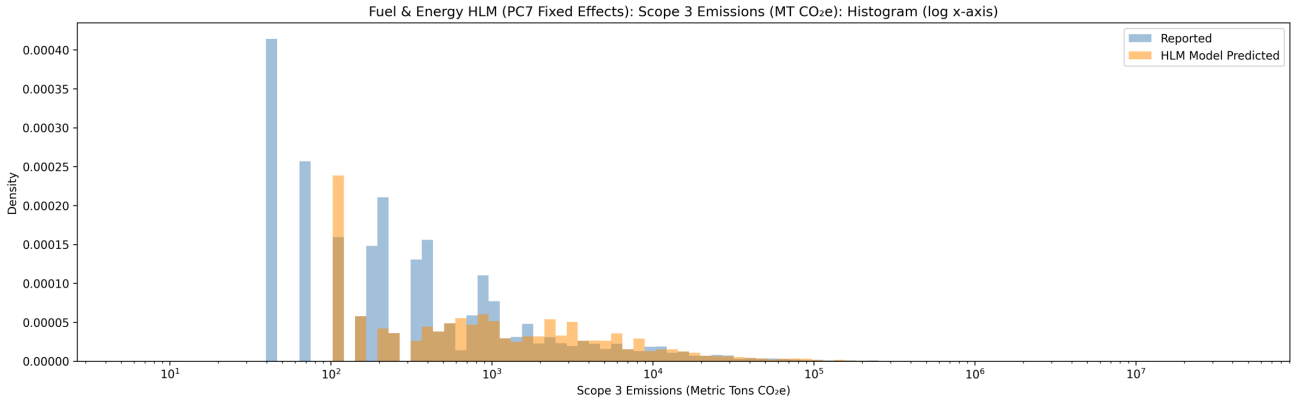


PC7 HLM Results



Conditional R^2: 87.81%
Marginal R^2: 74.46%
MAPE: 334.41%
MAPE(log): 9.51%
MAE(log): 0.90
RMSLE: 1.24
RMSE: 2571516.39
MAE: 508208.51





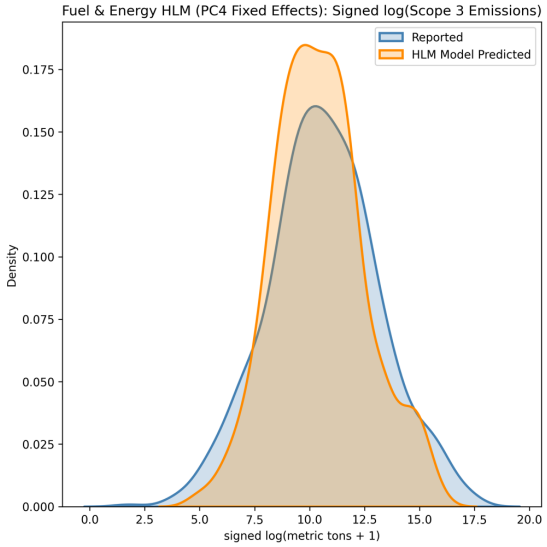
Fuel & Energy HLM with PC7 Fixed Effects: Test Set Performance Metrics:

Conditional R^2:	87.81%
Marginal R^2:	74.46%
MAPE:	334.41%
MAPE(log):	9.51%
MAE(log):	0.90
RMSLE:	1.24
RMSE:	2571516.39
MAE:	508208.51

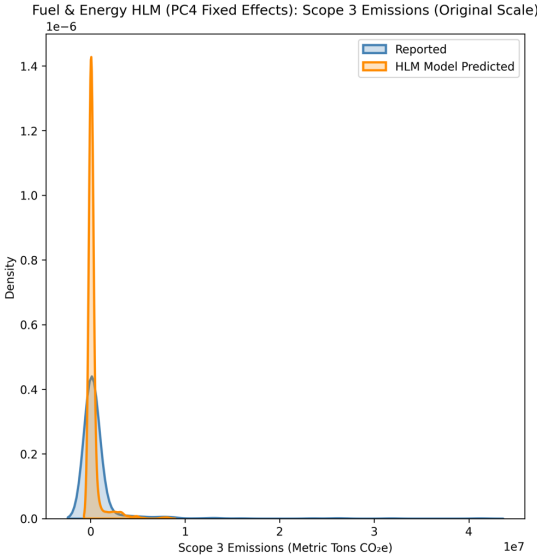
Top Influential Features (by p-value):

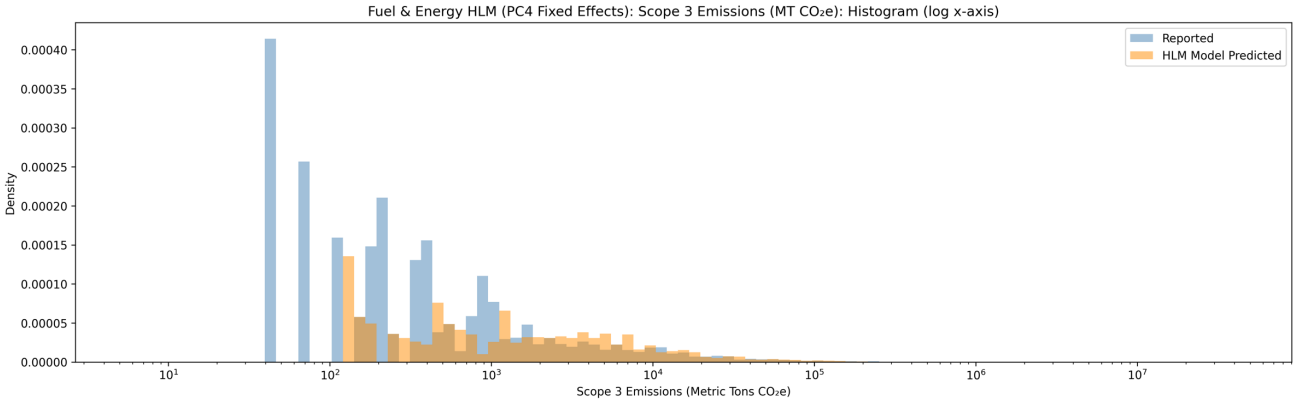
	coef	pvalue
PC7_1	0.409003	4.658019e-63
PC7_3	-0.335608	8.268940e-40
PC7_4	0.578541	2.700946e-38
Group Var	1.095801	5.374887e-37
PC7_6	0.367821	4.024290e-22
PC7_5	0.267187	1.083209e-14
C(Primary_activity)[T.Electricity networks]	3.537052	1.430333e-10
C(Primary_activity)[T.CCGT generation]	3.221812	3.652580e-09
C(Countries)[T.Brazil]	-2.806214	3.120196e-08
C(Primary_activity)[T.Gas utilities]	3.535807	1.884628e-07

PC4 HLM Results



Conditional R²: 88.28%
Marginal R²: 72.03%
MAPE: 461.44%
MAPE(log): 10.31%
MAE(log): 0.98
RMSLE: 1.35
RMSE: 2557300.68
MAE: 514015.59





Fuel & Energy HLM with PC4 Fixed Effects: Test Set Performance Metrics:

Conditional R²:	88.31%
Marginal R²:	71.93%
MAPE:	473.77%
MAPE(log):	10.36%
MAE(log):	0.98
RMSLE:	1.35
RMSE:	2561281.50
MAE:	515717.13

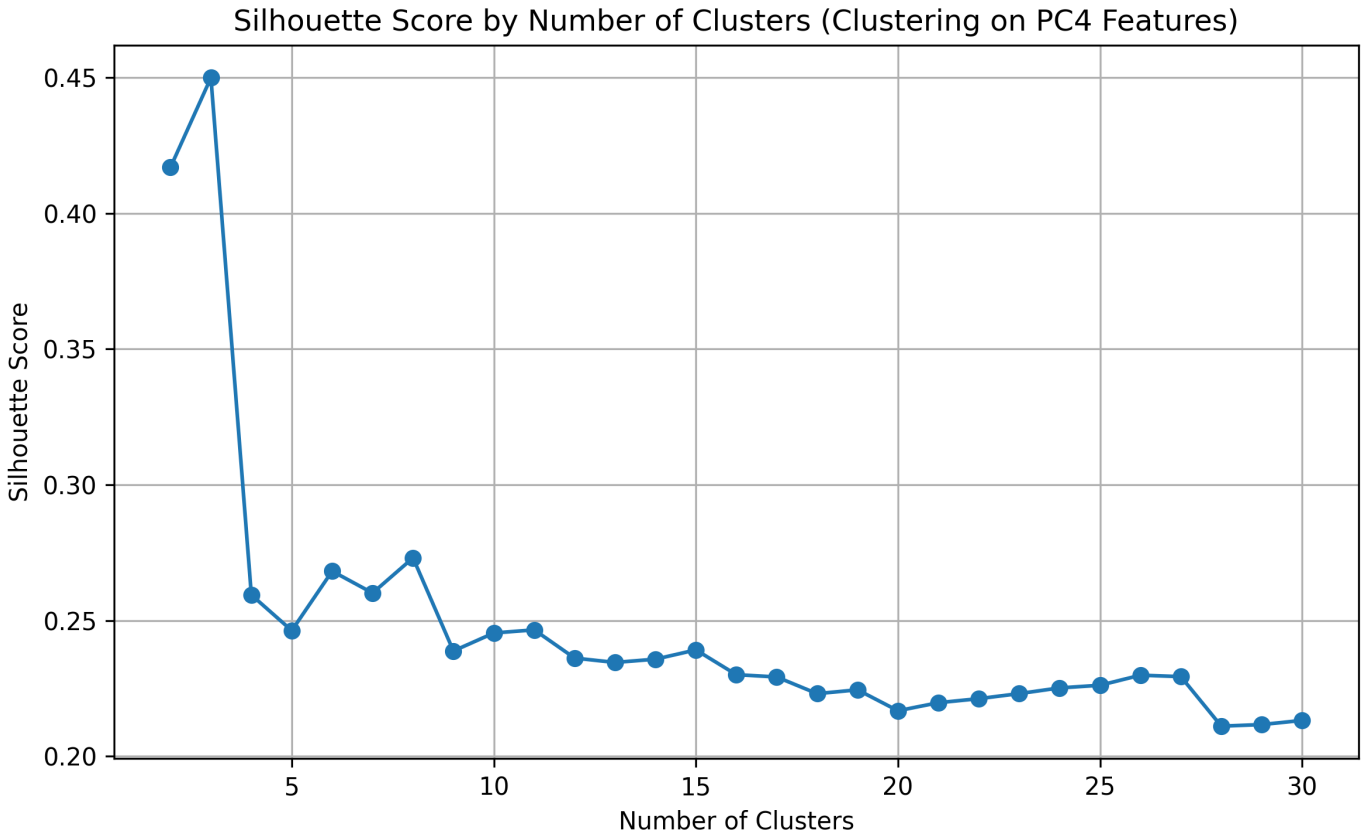
Top Influential Features (by p-value):

	coef	pvalue
PC4_1	0.377284	7.388114e-48
Group Var	1.400979	3.402177e-41
PC4_3	-0.259490	6.209189e-25
PC4_4	0.344483	2.953719e-17
C(Primary_activity)[T.Electricity networks]	4.436926	1.393702e-13
C(Primary_activity)[T.CCGT generation]	4.264294	5.763553e-13
C(Primary_activity)[T.Passenger airlines]	3.941496	1.272716e-10
C(Primary_activity)[T.Gas utilities]	4.163763	2.372617e-08
C(Primary_activity)[T.Inorganic base chemicals]	3.848034	1.352269e-07
C(Primary_activity)[T.Cement]	4.109366	6.009759e-07

Hierarchical Clustering on Principal Components

- Introducing hierarchical clustering on principal components
- Groups companies by similar patterns in first four principal components
- Aims to uncover hidden structure in the data
- Cluster assignments used as new features in HLM models
- Expected to improve prediction accuracy by capturing group-level differences

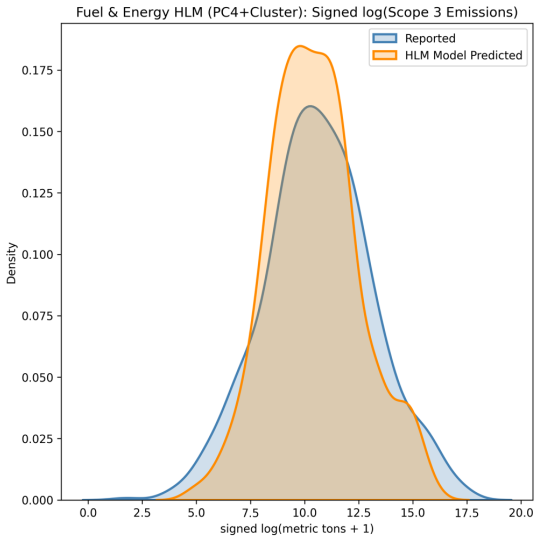
Clustering Silhouette Score with PC4



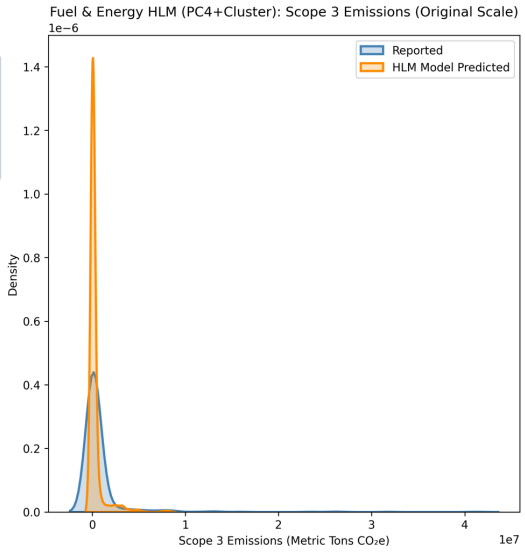
Number of observations in each cluster:

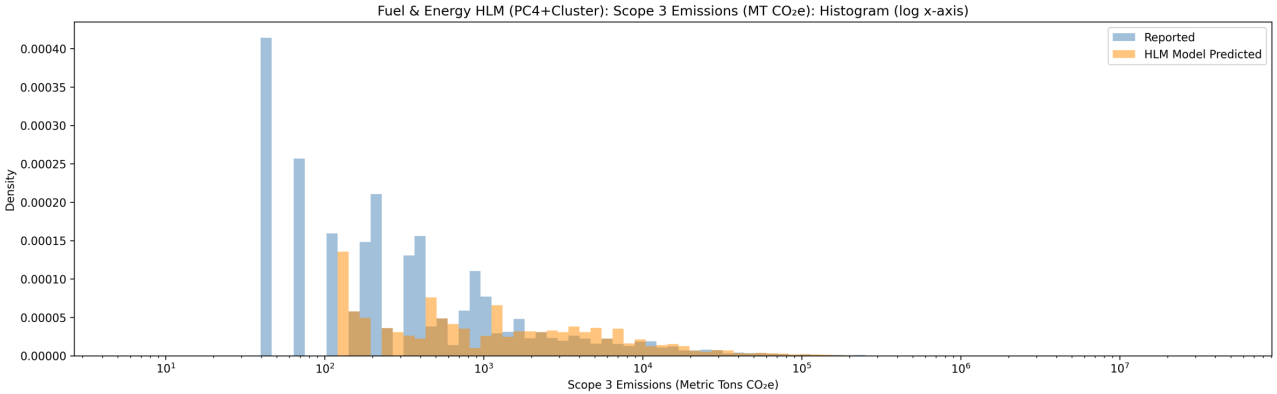
- 1st cluster - 369 observations
- 2nd cluster - 126 observations
- 3rd cluster - 3492 observations

PC4+Cluster HLM Results



Conditional R^2: 88.28%
Marginal R^2: 72.03%
MAPE: 461.44%
MAPE(log): 10.31%
MAE(log): 0.98
RMSLE: 1.35
RMSE: 2557300.68
MAE: 514015.59





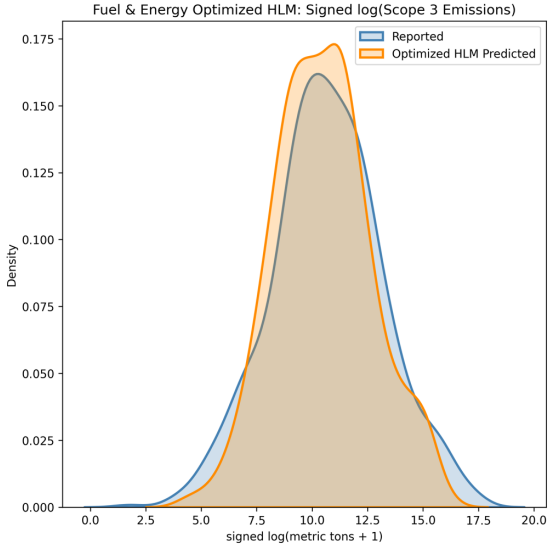
Fuel & Energy HLM with PC4+Cluster Fixed Effects: Test Set Performance Metrics:

Conditional R^2:	88.28%
Marginal R^2:	72.03%
MAPE:	461.44%
MAPE(log):	10.31%
MAE(log):	0.98
RMSLE:	1.35
RMSE:	2557300.68
MAE:	514015.59

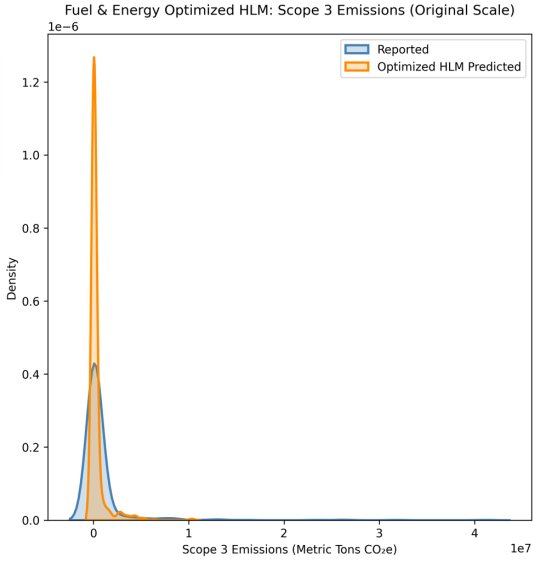
Top Influential Features (by p-value):

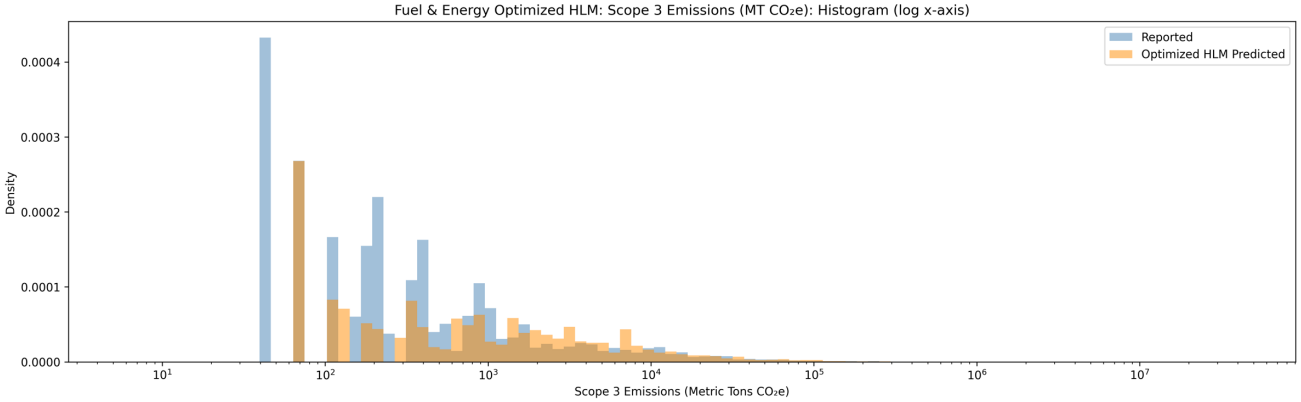
	coef	pvalue
PC4_1	0.385280	9.658610e−49
Group Var	1.387015	4.450913e−41
PC4_4	0.383662	1.616559e−17
C(Primary_activity)[T.Electricity networks]	4.395302	2.165646e−13
C(Primary_activity)[T.CCGT generation]	4.199226	1.329084e−12
C(Primary_activity)[T.Passenger airlines]	3.901442	1.898207e−10
PC4_3	−0.209330	7.040696e−09
C(Primary_activity)[T.Gas utilities]	4.122768	3.079677e−08
C(Primary_activity)[T.Inorganic base chemicals]	3.800880	1.848115e−07
C(Countries)[T.Brazil]	−2.739164	8.082591e−07

Fuel & Energy Optimized HLM Model Results



Conditional R^2: 87.62%
Marginal R^2: 76.12%
MAPE: 164.79%
MAPE(log): 8.40%
MAE(log): 0.80
RMSLE: 1.11
RMSE: 2545061.38
MAE: 477732.56





Optimized Mixed-Effects Model: Test Set Performance Metrics:

Conditional R^2:	87.62%
Marginal R^2:	76.12%
MAPE:	164.79%
MAPE(log):	8.40%
MAE(log):	0.80
RMSLE:	1.11
RMSE:	2545061.38
MAE:	477732.56

Top Influential Features (by p-value):

	coef	pvalue
slog_scope_1_2_emissions_amount	0.505167	1.793366e-79
Group Var	0.928626	1.213769e-36
C(Primary_activity)[T.Electricity networks]	3.001944	8.965375e-09
C(Countries)[T.Brazil]	-2.500436	1.391396e-07
C(Primary_activity)[T.Gas utilities]	3.264908	2.919022e-07
C(Primary_activity)[T.CCGT generation]	2.597995	5.494161e-07
C(Primary_activity)[T.Passenger airlines]	2.578735	1.427416e-06
C(Countries)[T.China]	-2.167157	1.551757e-06
C(Primary_activity)[T.Semiconductors]	3.312361	3.335800e-06
C(Primary_activity)[T.Other food processing]	2.105063	6.547823e-05

Fuel & Energy Model Metrics Summary

Model	PC Explained Variance	Conditional_R2	Marginal_R2	MAE_log	RMSLE
HLM PC7	81.88%	87.81%	74.46%	0.90	1.24
HLM PC4	62.43%	**88.31%**	71.93%	0.98	1.35
HLM PC4+Cluster	62.43%	88.28%	72.03%	0.98	1.35
Optimized Fuel Energy HLM		87.62%	**76.12%**	**0.80**	**1.11**

Best values highlighted. All models trained and tested on Fuel & Energy Scope 3 data.

Conclusion & Future Work

- Most Scope 3 emissions variance is explained by a few key features.
- The optimized HLM model was consistently accurate and reliable across all sources.
- Simpler models outperformed more complex PCA and clustering approaches.
- Future Work: The optimized HLM models can be retrained as new data becomes available to accommodate new information and boost accuracy.
- These results provide a strong foundation for transparent and reliable Scope 3 estimation that can evolve with new data.

Acknowledgments / Q&A

Thanks to the GHG Emissions Research group, Dr. Sorauf and Dr. Polson for the opportunity to work on this project

Questions?

References

[1] Standards & Guidance | GHG Protocol. [Online]. Available: https://ghgprotocol.org/standards-guidance. [Accessed: 19-Aug-2025].

[2] Financial Modeling Prep - FinancialModelingPrep. [Online]. Available: https://site.financialmodelingprep.com/. [Accessed: 28-Jun-2025].

[3] G. Serafeim and G. Velez Caicedo, "Machine learning models for prediction of scope 3 carbon emissions," Harvard Business School Accounting & Management Unit Working Paper, no. 22–080, 2022.

[4] Q. Nguyen, I. Diaz-Rainey, A. Kitto, B. I. McNeil, N. A. Pittman, and R. Zhang, "Scope 3 emissions: Data quality and machine learning prediction accuracy," PLoS Climate, vol. 2, no. 11, p. e0000208, 2023.

[5] B. Ebeling, C. Vargas, and S. Hubo, "Combined cluster analysis and principal component analysis to reduce data complexity for exhaust air purification," Open Food Sci J, vol. 7, pp. 8–22, 2013.

[6] H. Abdi and L. J. Williams, "Principal component analysis," Wiley Interdisciplinary Reviews: Computational Statistics, vol. 2, no. 4, pp. 433–459, 2010.

Supplementary Material

Scope 3 Source Type	Train Observations	Test Observations	Total Number of Observations	HLM (MAE_log)	Industry Fill (MAE_log)	Naïve OLS (MAE_log)	Full OLS (MAE_log)	Stepwise OLS (MAE_log)	Linear Forest (MAE_log)
Business Travel	4130	789	4919	0.83	0.96	0.99	0.94	0.94	0.94
Purchased Goods and Services	3483	724	4207	1.18	1.71	1.72	1.73	1.69	1.74
Waste Generated in Operations	3282	678	3960	1.13	1.44	1.45	1.47	1.42	1.40
Fuel and Energy	3242	679	3921	0.80	1.44	1.48	1.46	1.52	1.41
Employee Commuting	3092	653	3745	0.84	1.22	1.18	1.17	1.11	1.11
Upstream Transportation and Distribution	2643	555	3198	1.20	1.56	1.48	1.52	1.47	1.48
Capital Goods	2218	505	2723	1.01	1.42	1.27	1.24	1.15	1.25
Downstream Transportation and Distribution	1630	303	1933	1.45	1.55	1.50	1.51	1.43	1.47
Use of Sold Products	1568	336	1904	1.26	1.63	1.63	1.72	1.61	1.66
End of Life Treatment of Sold Products	1537	342	1879	1.33	1.81	1.71	2.05	1.61	1.66
Upstream Leased Assets	779	158	937	1.38	2.27	1.78	1.98	1.66	1.75
Downstream Leased Assets	749	159	908	1.34	2.65	2.08	2.46	1.91	2.04
Investments	572	120	692	1.56	3.12	2.30	2.76	1.87	2.27
Processing of Sold Products	429	91	520	1.25	3.10	2.44	2.47	2.38	2.55
Franchises	259	56	315	1.03	2.47	1.95	3.59	1.60	2.20
Other (Upstream)	169	18	187	1.30					
Other (Downstream)	58	12	70	1.09					

Supplementary Table 1. MAE-log comparison across models [4]

Supplementary Material

Scope 3 Source Type	Train Observations	Test Observations	Total Number of Observations	HLM (RMSLE)	OLS (RMSLE)	GLM (RMSLE)	KNN (RMSLE)	RF (RMSLE)	AdaBoost (RMSLE)
Business Travel	4130	789	4919	1.10	1.33	1.36	1.08	0.98	0.90
Purchased Goods and Services	3483	724	4207	1.67	2.68	2.70	2.06	1.88	1.66
Waste Generated in Operations	3282	678	3960	1.43	1.98	2.05	1.57	1.49	1.42
Fuel and Energy	3242	679	3921	1.11	1.88	1.91	1.51	1.45	1.34
Employee Commuting	3092	653	3745	1.15	1.73	1.75	1.37	1.24	1.19
Upstream Transportation and Distribution	2643	555	3198	1.61	2.11	2.13	1.77	1.48	1.37
Capital Goods	2218	505	2723	1.38	2.23	2.21	1.69	1.49	1.39
Downstream Transportation and Distribution	1630	303	1933	1.91	2.36	2.31	1.83	1.66	1.55
Use of Sold Products	1568	336	1904	1.65	2.66	2.69	1.84	1.65	1.36
End of Life Treatment of Sold Products	1537	342	1879	1.71	2.74	2.85	2.32	1.97	1.81
Upstream Leased Assets	779	158	937	1.77	2.25	2.32	2.18	1.84	1.87
Downstream Leased Assets	749	159	908	1.92	2.59	2.69	2.16	1.91	1.76
Investments	572	120	692	2.01	2.42	2.43	1.69	1.49	1.35
Processing of Sold Products	429	91	520	1.82	2.77	2.71	2.07	1.74	1.59
Franchises	259	56	315	1.40	2.57	2.82	2.10	1.55	1.18
Other (Upstream)	169	18	187	2.47					
Other (Downstream)	58	12	70	1.40					

Supplementary Table 2. RMSLE comparison across models [3]