

UC Irvine

UC Irvine Electronic Theses and Dissertations

Title

Making Garbage Great Again: Classification and network-based emissions modeling of California's waste management and recycling system

Permalink

<https://escholarship.org/uc/item/7ts0f7m5>

ISBN

9798252413600

Author

Casebolt, Brian Thomas

Publication Date

2026-06-15

Copyright Information

This work is made available under the terms of a Creative Commons Attribution-NonCommercial-ShareAlike License, available at <https://creativecommons.org/licenses/by-nc-sa/4.0/>

Peer reviewed|Thesis/dissertation

UNIVERSITY OF CALIFORNIA,
IRVINE

Making Garbage Great Again: Classification and Network-Based Emissions Modeling of
California's Waste Management and Recycling System

DISSERTATION

submitted in partial satisfaction of the requirements
for the degree of

DOCTOR OF PHILOSOPHY

in Civil and Environmental Engineering

by

Brian Thomas Casebolt

Dissertation Committee:
Professor Jean-Daniel Saphores, Chair
Associate Professor Michael Hyland
Professor Michael McNally

2026

DEDICATION

To

my wife
my children
my parents

in recognition of their love and support

TABLE OF CONTENTS

DEDICATION	ii
TABLE OF CONTENTS.....	iii
LIST OF FIGURES	vi
LIST OF TABLES	x
ACKNOWLEDGMENTS	xii
VITA.....	xiv
ABSTRACT OF THE DISSERTATION	xv
Chapter 1 Introduction	1
Chapter 2 Background	10
2.1 Institutional and Regulatory Framework	10
2.2 Data Components of California’s Regulatory Framework	15
2.3 Material Classes and Processing Pathways.....	23
2.4 Waste Management Infrastructure and Facility Types	25
2.5 Empirical Material-Facility Interactions.....	36
2.6 Limitations of Material and Facility Typologies	44
Chapter 3 Machine learning approach to identifying recycling and waste management facilities based on their inputs and outputs.....	48
3.1 Introduction.....	48
3.2 Literature Review: Classification and Integrated Waste Management Systems	52
3.3 Data Description and Pattern Specification	59
3.4 Classification Methodology	62
3.5 Implementation	75

3.6 Cluster Descriptions.....	93
3.7 Outcomes	102
3.8 Conclusions.....	106
Chapter 4 Quantifying the Carbon Footprint of California’s Waste Management and Recycling Operations.....	108
4.1 Introduction.....	108
4.2 Literature Review: Emissions Modeling Across Waste Management Systems	110
4.3 Methods.....	119
4.4 Synthesis	165
Chapter 5 Conclusions	184
5.1 Principal Findings and Conclusions	186
5.2 Policy Implications for California Waste Management and Climate Planning	187
5.3 Future Research	189
References.....	190
Appendix A – Material Dendrogram	191
Appendix B – Intermediate Calculations.....	194
B.1 Survey of vehicle fleet activity.....	194
B.2 System Partitions.....	195
Appendix C – Calibrating Correspondence Analysis	197
C.1 Iteration 0 (n=1686)	197
C.2 Iteration 1 (1568).....	198
C.3 Iteration 2 (n=1398)	200
C.4 Iteration 3 (n=1329)	201

C.5 Column Standard Coordinates for the first 17 principal dimensions:	205
C.6 Inertias and Eigen Values:.....	206
Appendix D – Miscellaneous Methods.....	207
D.1 Isometric Log Ratio (ILR) Model.....	208
D.2 MRF Material Flow Model.....	210
Appendix E – Emissions Modeling and EMFAC.....	215
E.1 On-Road Model implementation.....	217

LIST OF FIGURES

Figure 2.1 Map of Transfer Stations and Processors	27
Figure 2.2 Map of MRF locations.....	29
Figure 2.3 Map of Recyclers and Composters.....	31
Figure 2.4 Map of Landfills	34
Figure 2.5 Histogram of Total Reported Outflows from Transfer/Processor Activities	39
Figure 2.6 Histogram of Total Reported Inflows to Landfill Activities	40
Figure 2.7 Histogram of Total Reported Outflows from Recycler/Composter Activities.....	41
Figure 2.8 Box Plots of Facility-level Nonzero Inflows and Outflows	46
Figure 3.1 General pattern recognition paradigm	49
Figure 3.2 Biplot of Activity and Material Alignment in 2 Dimensions (Unit Circle)	75
Figure 3.3 Scree Plot of Feature Dimension Eigen-values	76
Figure 3.4 Cumulative Inertia Chart and selecting dimensionality reduction	78
Figure 3.5 Inertia and Silhouette Score Charts for K-means screening runs.....	84
Figure 3.6 Dendrogram for Hierarchical Clustering.....	87
Figure 3.7 Silhouette Score and Linkage distance Charts for Hierarchical screening runs.....	88
Figure 3.8 Screening of bandwidth quantile parameter for Mean Shift clustering.....	89
Figure 3.9 Eigengap chart for choosing k in spectral clustering.....	91

Figure 3.10 Heatmap of coassociation matrix, sorted by consensus cluster label	93
Figure 3.11 Gross inflow and outflow balances for processors/recyclers	98
Figure 3.12 Compositional profile of Cluster 6	98
Figure 3.13 Compositional profiles of clusters 0 and 5	100
Figure 3.14 Biplot of Reclassified Activity and Material Alignment in 2 Dimensions (Unit Circle)	104
Figure 4.1 Graph Representation of Paths, Pathsegments, and TAZ (CoAbDis).....	129
Figure 4.2 PMF of Heavy Vehicle Population Age (application year 2021).....	138
Figure 4.3 EMFAC Off-Road deterioration factors.....	139
Figure 4.4 Nonlinearity (Economies of Scale) in Landfill Heavy Equipment Usage	146
Figure 4.5 Block Diagram of MRF Process Flow	152
Figure 4.6 Feasible Recovery and Inflow Composition Sub-Space	155
Figure 4.7 Material Stream Purity Estimates	159
Figure 4.8 Baler Power Output Profiles per Cycle	163
Figure 4.9 On-Road GHG Intensity Comparison	167
Figure 4.10 Distribution of NO _x and PM for On-Road Activity.....	168
Figure 4.11 On-Site GHG Emissions Inventory	169
Figure 4.12 On-Site NO _x Emissions Inventory	169

Figure 4.13 On-Site PM Emissions Inventory (Metric Tons)	170
Figure 4.14 Percentage Contribution of Waste Management Activities to Total GHG	171
Figure 4.15 Percentage Contribution of Waste Management Activities to Total PM	171
Figure 4.16 Percentage Contribution of Waste Management Activities to Total NOx	172
Figure 4.17 Percentage Contribution of Waste Management Activities to Total Fuel Usage....	172
Figure 4.18 Miles per ton by material stream	177
Figure 4.19 Sankey Diagram of Carbon Emissions by Destination	178
Figure 4.20 Solid Waste to Disposal Path	180
Figure A.0.1 Material Hierarchy (Mixed Waste Branch)	191
Figure A.0.2 Material Hierarchy (Other Waste Branch)	192
Figure A.0.3 Material Hierarchy (Mixed Recyclables Branch).....	193
Figure C.0.1 Biplot result for Iteration 0	198
Figure C.0.2 Biplot result for Iteration 1	199
Figure C.0.3 Biplot result for Iteration 2	200
Figure C.0.4 Biplot (three dimensional) for Iteration 2	201
Figure C.0.5 Biplot Result for Iteration 3	202
Figure C.0.6 Biplot result (three dimensional) for final iteration	203
Figure C.0.7 Distribution of Cosines	204

Figure E.0.1 Map of CoAbDis Subareas	216
Figure E.0.2 Default Speed Distributions for T7 Vehicle Categories	219

LIST OF TABLES

	Page
Table 2.1 Reporting Entity Type and Frequency	19
Table 2.2 Material Streams between Activities (2021 short tons).....	37
Table 2.3 Material Outflows ('21 short tons) by Facility Activity	42
Table 3.1 Material categories in RDRS and reported tons of 2021 outflows	60
Table 3.2 Equations to apply the estimated CoAn transform to Row- or Column-profiles	65
Table 3.3 Optimal feature reference vectors across 9 dimensions.....	82
Table 3.4 Intra- and Inter-Cluster Coassociation.....	92
Table 3.5 Composition and size profiles of clusters	94
Table 3.6 Summary of flow contributions by classification	103
Table 3.7 Summary of emissions models by classification	106
Table 4.1 EMFAC On-road specification summary to retrieve T7 SWCV Inventory	124
Table 4.2 EMFAC On-road query specification summary to retrieve T7 Tractor Emission Rates	132
Table 4.3 Emissions Tier Base Year Population	137
Table 4.4 Heavy Equipment Load Factor	141
Table 4.5 Annual horsepower hours and tons disposed for respondent landfills	144

Table 4.6 Landfill Heavy Vehicle Specific Energy Model Significance.....	145
Table 4.7 Operational Purity Values.....	158
Table 4.8 MRF Separation Process Specific Energy and Emissions.....	161
Table 4.9 Baling Emissions	164
Table 4.10 Statewide Carbon Footprint by Source in Base Year ('21)	166
Table 4.11 Brokerage quantity by material.....	174
Table 4.12 Tons of material per origin, destination facility types	175
Table 4.13 VMT per ton by material stream.....	177
Table 4.14 Disposal Subsystem Results	182
Table B.0.1 Partition Schema	196
Table C.0.1 Pre-CoAn exclusion conditions.....	202
Table C.0.2 Column Standard Coordinates for 17 dimensions	205
Table C.0.3 Eigen values and contribution to inertia.....	206
Table E.0.1 Statewide Vehicle Type 7 (GVWR8) and Fuel Distribution, Calendar Year 2021	218

ACKNOWLEDGMENTS

This dissertation would not have been possible without the sustained support, guidance, and intellectual leadership of my advisor, Dr. Jean-Daniel Saphores, whose mentorship shaped every stage of this work. As principal investigator of the REMADE project, he provided both the research framework and the critical feedback that guided the development, refinement, and completion of this dissertation. I am especially grateful for the considerable time he devoted to my work, and for his insistence on clarity, discipline, and technical depth. I also wish to acknowledge the important contribution of Dr. Craig Rindt, whose work on the REMADE project “Pipeline” was foundational to the vehicle routing components of Chapter 4. I am also grateful to Professors Ritchie, Recker, Tok, McNally, and Hyland for their instruction, encouragement, and thoughtful feedback across coursework, research, and professional development over the years, which collectively strengthened both the technical rigor and applied relevance of this work.

I am also appreciative of the collaboration and cooperation provided by colleagues in public agencies and industry. Dan Brown and his data engineering colleagues at CalRecycle were exceptionally generous with their time, expertise, and data access, which were foundational to both the REMADE project and this dissertation. I also thank the managers and employees across numerous waste handling and recycling operations who answered questions, provided site access, and shared operational insight that grounded this research in real-world practice. In particular, I acknowledge David Wong and Jack Du at Orange County Waste & Recycling; Saul Avila, Jose Talavera, Claudia Eguihua, Ricardo Padilla, and Wyatt Anderson at Waste Management; Risa Baron at Republic Services; Lisa Thompson, Carlos Guerrero, and Kathryn

Aguilar at Riverside County Department of Waste Resources; Ted Smith at Potential Industries Inc.; and Dannette Fieguth at Chicago Grade. Their willingness to engage openly with researchers made this work both feasible and substantially stronger.

Financial support for this research was provided by multiple sources. I gratefully acknowledge fellowship support from the Pacific Southwest Region University Transportation Center (PSR UTC), which supported me across six years of my education, and enabled the development and application of new technical skills, particularly in computational analysis and programming. This dissertation was directly supported by the U.S. Department of Energy through the REMADE Institute under Project 21-01-SA-5104. The views and conclusions expressed in this dissertation are those of the author and do not necessarily reflect those of the funding agencies.

VITA

Brian Thomas Casebolt

EDUCATION

Ph.D., Transportation Systems Engineering, 2026
University of California Irvine

M.S., Transportation Systems Engineering, 2018
University of California Irvine

Lean Six Sigma Certification, 2015
University of Michigan Engineering, Online & Professional Education

B.S., Chemical Engineering, 2012
University of Southern California

FIELD OF STUDY

Interdisciplinary: Transportation Economics and Planning
ITS - Transportation Science (2015 – 2017)

Civil and Environmental Engineering
ITS - Transportation Systems Engineering (2017 – 2026)

HONORS AND AWARDS

PSR-UTC Fellowship Recipient 2017, 2018, 2019, 2020, 2021, 2022

TEACHING EXPERIENCE

Teaching Assistantships:

CEE111: Systems Analysis and Decision Making (Winter '18)

CEE121: Transportation Systems I- Analysis and Design (Fall '18, '19, '20, '21)

CEE122: Transportation Systems II- Operations and Control (Spring '17)

CEE123: Transportation Systems III- Planning and Forecasting (Spring '25)

GRADUATE STUDENT RESEARCH PROJECTS

REMADE Project (2023 – 2025)

CARB Vocational Aerodynamics Project (2018)

PROFESSIONAL AFFILIATIONS

ITE 2026

OCTEC 2026

ABSTRACT OF THE DISSERTATION

Making Garbage Great Again: Classification and Network-Based Emissions Modeling of
California's Waste Management and Recycling System

by

Brian Casebolt

Doctor of Philosophy in Civil and Environmental Engineering

University of California, Irvine, 2026

Professor Jean-Daniel Saphores, Chair

Waste management systems are increasingly expected to satisfy environmental, climate, and resource recovery objectives, yet engineering analyses remain constrained by limited empirical characterization of facility operations and infrastructure networks. Although regulatory databases contain extensive information on material flows, they generally do not report the operational activity responsible for energy consumption and emissions, forcing many inventories to rely on representative facilities or generalized emission factors. This dissertation develops an engineering framework for estimating the energy consumption and emissions of California's post-collection waste management system by integrating empirical facility classification, statewide transportation network modeling, and facility activity estimation.

The analysis demonstrates that administrative facility categories frequently obscure substantial operational heterogeneity and therefore provide an inadequate basis for engineering emissions modeling. A flow-based classification derived from observed material inflows and outflows produces more representative facility archetypes for assigning process parameters and estimating activity. Using these classifications together with statewide routing analysis and facility-specific operating characteristics, this dissertation develops California-specific estimates of transportation activity, fuel consumption, energy use, greenhouse gas emissions, and criteria pollutants across major components of the post-collection waste management system.

The results further demonstrate that emissions inventories depend critically on infrastructure configuration. Transportation demand, equipment activity, and processing requirements emerge from the interaction of facility operations, material routing, and network structure rather than from fixed assumptions regarding transport distances or representative facilities. Consequently, meaningful evaluation of waste management policies requires integrated consideration of transportation, processing, and disposal rather than analysis of individual components in isolation.

Beyond producing a statewide emissions inventory, this dissertation establishes an empirical framework for transforming incomplete administrative records into defensible engineering estimates of system activity. The resulting methodology and inventory provide a baseline for future infrastructure planning, climate policy evaluation, and analysis of evolving waste management strategies in California.

Chapter 1 Introduction

Waste management has emerged as one of the defining environmental, infrastructural, and policy challenges of the twenty-first century. Globally, rising material consumption, urbanization, and population growth have driven steady increases in municipal solid waste (MSW) generation, while the material composition of that waste stream has become more heterogeneous and operationally complex. At the same time, traditional disposal-oriented paradigms—centered on landfilling and long-distance hauling—are increasingly constrained by land scarcity, public opposition, climate objectives, and regulatory mandates. These pressures have shifted policy attention toward waste reduction, recycling, organics diversion, and circular-economy strategies, all of which depend critically on the performance of post-collection waste management and recycling infrastructure.

California state law combines ambitious statutory diversion and climate targets with one of the largest, most complex waste management systems in the world. Policies such as Senate Bill 1383 (organic waste diversion and methane reduction), Senate Bill 54 (plastic pollution reduction and extended producer responsibility), and long-standing landfill diversion mandates have fundamentally altered the incentives, flows, and technologies that define California’s waste system (CalRecycle, 2026a, 2025). Yet despite this policy sophistication, core elements of the system remain poorly characterized from an engineering and systems perspective. In particular, there is limited empirical understanding of how facilities function within the integrated network of collection, transfer, processing, and disposal and how those functions translate into energy use, transportation activity, and emissions.

At the heart of the system are waste management facilities: transfer stations, material recovery facilities (MRFs), composting and anaerobic digestion facilities, construction and demolition processors, landfills, and specialized recyclers. These facilities are not merely endpoints; they are active nodes in a highly interconnected network that determines material routing, vehicle activity, equipment use, and ultimately environmental impacts. Accurately modeling system performance therefore requires a representation of facilities that reflect their operational roles beyond regulatory labels.

Existing administrative classifications, such as those embedded in CalRecycle's Recycling and Disposal Reporting System (RDRS) and permitting databases, were developed primarily for compliance and reporting purposes. As a result, they often collapse heterogeneous operations into broad categories (e.g., "Transfer/Processor" or "Recycler/Composter") that obscure meaningful differences in material handling, throughput, and processing intensity. Even permit data, when available, does not accurately describe throughput for the full scope of processes and materials. This mismatch between administrative categories and operational reality constitutes a fundamental problem: without a defensible, data-driven understanding of facility behavior, system-level models of waste flows, transportation demand, and emissions risk systematic bias.

From an engineering standpoint, two interrelated gaps motivate this dissertation. First, there is a lack of empirical, system-scale facility classification frameworks that derive operational archetypes directly from observed material flows. While machine-learning methods have been widely applied to waste sorting and composition at the object or image level (Fotovvatikhah et al., 2025; Yevle and Mann, 2025), there are no empirical clustering/classification applications at the facility level. This is presumably due to lack of data

availability at scale, because without regulatory incentives, there is no reason that operators would share this information. In fact, there was and continues to be a great deal of stakeholder negotiation and pushback as CalRecycle develops rules of implementation (CalRecycle, 2019, 2013).

Second, emissions and energy modeling of waste management systems frequently proceeds independently of facility classification. Transportation emissions are often estimated using aggregate vehicle miles traveled or simplified assumptions about routing (U.S. Environmental Protection Agency, 2016; Vergara et al., 2011), while on-site emissions are approximated using generic per-ton factors with questionable temporal and geographic transferability (IPCC, n.d.; U.S. Environmental Protection Agency, 2016). These approaches rarely distinguish between different types of transfer stations, processors, or recyclers, even though their equipment usage, duty cycles, and throughput differ substantially. Consequently, emissions inventories may fail to capture spatial heterogeneity, operational scale effects, and the coupling between facility function and transport demand.

This dissertation addresses a broader methodological challenge in waste systems engineering: developing empirically grounded, system-level representations of post-collection waste infrastructure capable of supporting defensible estimates of energy consumption and greenhouse gas emissions. Existing studies frequently examine individual facilities or evaluate predefined management scenarios, but large-scale analyses often rely on stylized assumptions regarding transport distances, facility operations, and recovery efficiencies. Transportation distances are typically treated as fixed scenario inputs rather than emergent outcomes of the underlying infrastructure network, while process-specific emissions are commonly represented

using generalized factors that are only loosely coupled to the operational characteristics of individual facilities.

These limitations are particularly important because emissions from waste management systems are fundamentally determined by the interaction between activity and emission intensity. Transportation emissions depend on ton-kilometers traveled, vehicle class, routing, and infrastructure configuration, while operational emissions depend on the technologies and processes employed at individual facilities. Consequently, assumptions regarding facility function and network configuration directly influence estimated environmental outcomes. Administrative classifications, however, often do not correspond to actual operational behavior, creating uncertainty regarding which models and parameters should be applied.

Accordingly, this dissertation develops an integrated framework that combines facility classification, empirically observed material flows, transportation network modeling, and operational emissions estimation within a unified analytical framework. Rather than assuming stylized transport distances or fixed recovery efficiencies, it treats energy use and carbon emissions associated with household waste collection, recycling, and disposal as emergent properties of empirically observed infrastructure configurations and operational parameters. The resulting framework provides a baseline representation of California's waste management system while establishing a methodology that can be extended to evaluate future policy scenarios and infrastructure changes.

The core engineering problem addressed in this dissertation is therefore how to construct an empirically grounded, reproducible representation of California's waste management system that simultaneously classifies facilities based on their observed material handling, embeds those classifications within a statewide, network-based representation of waste transportation, and

supports estimation of transportation, operational, and process-based emissions across facility types. Addressing this problem requires integrating methods from machine learning, compositional data analysis, network modeling, and emissions estimation—applied to large, imperfect regulatory datasets.

The overarching purpose of this dissertation is to develop and apply a system-level, data-driven framework for characterizing waste management facilities and quantifying the carbon and air-pollutant impacts of California’s post-collection waste system. This purpose is pursued through three major research components, each addressing a distinct but interconnected set of questions. The first component focuses on facility classification. The guiding research questions are 1) “How to classify waste management facilities into stable, interpretable operational archetypes using reported inflow and outflow material data,” and 2) “How to use that information to refine emissions profiles?” the term profile is used here to mean the absolute amount of emissions and relative proportions from different sources or activities (composition and basis). The second component addresses transportation and operational emissions. Its primary focus is the magnitude and spatial distribution of greenhouse gas and criteria pollutant emissions associated with post-collection waste transport in California and how they differ across facility archetypes for subsectors of the waste management infrastructure. Together, these questions frame the dissertation as an integrated engineering analysis of infrastructure classification, transport networks, and emissions.

The subject of this dissertation is situated within broader global and policy trends related to resource efficiency, climate mitigation, and environmental justice. Worldwide, waste generation is projected to increase substantially over coming decades, particularly in rapidly urbanizing regions (Kaza et al., 2018). At the same time, international climate agreements and

national policies increasingly recognize waste management as a non-trivial source of greenhouse gas emissions, particularly methane from landfills and carbon dioxide from transportation and processing activities (Bruni et al., 2020; Bux et al., 2023; Vergara et al., 2011; Wang et al., 2020).

Global municipal solid waste (MSW) generation has increased dramatically over the past half century and is projected to continue rising through mid-century. Empirical and econometric analyses consistently find a positive elasticity between per-capita waste generation and income, with limited evidence of structural decoupling at higher development levels (Chen et al., 2020; Daskalopoulos et al., 1998). Even in high-income regions, aggregate waste generation continues to scale with GDP and consumption intensity.

Cross-national accounting frameworks further show that landfill remains a dominant global disposal pathway despite circular economy commitments (Tisserant et al., 2017). Waste-sector greenhouse gas emissions—particularly methane from landfills—are projected to increase substantially under business-as-usual scenarios (Chen et al., 2020). These global patterns suggest that improvements in diversion rates do not inherently translate into absolute reductions in environmental burdens.

Comparative performance studies across EU member states demonstrate substantial heterogeneity in diversion efficiency, even among economically advanced jurisdictions (Ginevičius, 2022; Ríos and Picazo-Tadeo, 2021). These studies highlight the importance of distinguishing between (1) waste generation intensity, (2) treatment pathway distribution, and (3) environmental performance metrics. Recycling rate alone is insufficient as a system performance indicator; boundary definition and treatment assumptions materially influence results.

Together, this global literature establishes two key implications for subnational modeling. First, waste systems must be evaluated using explicitly defined system boundaries that distinguish generation from treatment performance. Second, environmental assessment must incorporate transportation activity, processing intensity, and landfill emissions in an integrated framework. These principles guide the modeling approach developed in this dissertation.

In the United States, California has emerged as a policy leader by explicitly linking waste diversion strategies to climate objectives, particularly methane reduction and renewable energy generation from organic waste. These policy initiatives highlight the need to understand how diversion pathways interact with transportation distances, process energy requirements, and facility-level efficiency. In practice, however, those benefits are highly context-dependent and sensitive to transportation distances, processing energy requirements, and facility-level efficiency. Long-distance hauling to centralized facilities can erode methane-avoidance gains, while operational inefficiencies at material recovery facilities may increase residual disposal. Together, these dynamics highlight the need for system-level, spatially explicit modeling approaches that connect policy intent to the physical realities of waste management infrastructure.

California was selected as the study area for several reasons. First, the state maintains detailed historical-administrative datasets on waste flows through the Solid Waste Information System (SWIS) and Disposal Reporting System (DRS). In fact, recent legislation has resulted in a major evolution of DRS into the Recycling and Disposal Reporting System (RDRS), enabling facility-to-facility analysis at scale for all materials and waste streams. Second, California's regulatory environment has accelerated diversification of facility types and operational strategies, making it an ideal laboratory for studying classification and system behavior. Third,

the state's size and regional heterogeneity—from dense urban corridors to remote rural areas—allow examination of spatial variation in transport and emissions.

The modeling framework developed in this dissertation characterizes the post-collection waste management system, integrating inter-facility transportation and on-site processing activities within a unified emissions analysis. Residential curbside collection is incorporated through established emissions inventories, allowing the analysis to focus on the statewide infrastructure network that governs material transfer, processing, and disposal. The framework further establishes a baseline for system-level emissions accounting that provides context for evaluating landfill gas generation and future policy scenarios involving organics diversion. The analysis is based primarily on calendar year 2021, selected for data completeness and relevance to the current regulatory environment.

Several overarching assumptions were necessary. Reported RDRS flows were assumed to be correct but not mass-balanced, reflecting known limitations in reporting. Routing between facilities was approximated using shortest-path road networks. Emissions factors are derived from EMFAC and applied consistently across the modeled system. These assumptions are made explicit and evaluated for their implications in later chapters.

This dissertation makes several substantive contributions to the engineering analysis of waste management systems. First, it develops a novel facility classification framework based on ensemble-consensus clustering of material inflow and outflow profiles. By integrating correspondence analysis with multiple clustering algorithms, the approach produces stable, interpretable facility archetypes grounded in empirical behavior rather than administrative labels. Second, it embeds these classifications within a statewide, network-based representation of waste flows and transportation. This enables spatially resolved estimation of post-collection transport

activity and emissions, linking facility function directly to transportation demand. Third, it integrates transportation emissions, on-site equipment activity, and electricity usage into a unified carbon accounting framework. This integration allows consistent comparison of emissions sources and supports system-level evaluation of diversion and infrastructure strategies. Collectively, these contributions advance the methodological toolkit available for waste systems engineering and provide empirical insights based on available data that are relevant to policy evaluation and infrastructure planning.

The remainder of this dissertation is organized as follows. Chapter 2 provides background on California's waste management infrastructure and its regulatory framework. Chapter 3 presents the ensemble-consensus facility classification methodology and results. Chapter 4 develops the transportation and operational emissions framework for estimating the energy consumption and carbon footprint of household waste collection, recycling, and disposal, and presents statewide results. Chapter 5 provides conclusions and recommendations for future study and regulatory analysis.

Chapter 2 Background

This chapter provides the institutional, data, and operational background necessary to analyze California's post-collection waste management system. Section 2.1 introduces the institutional and regulatory framework governing waste management in California, including the evolution of the state's legislative approach to diversion, climate policy, and extended producer responsibility. Section 2.2 describes the principal regulatory databases used in this research, including the Solid Waste Information System (SWIS) and the Recycling and Disposal Reporting System (RDRS) and discusses their structure and limitations. Section 2.3 reviews the major material classes and their processing pathways, while Section 2.4 describes the principal facility types that comprise California's post-collection waste infrastructure. Section 2.5 then examines empirically observed material flows between these facilities, revealing the complexity of system interactions. Finally, Section 2.6 discusses the limitations of nominal material and facility classifications and motivates the facility classification framework developed in Chapter 3.

2.1 Institutional and Regulatory Framework

A coherent treatment of waste infrastructure in California must begin with the institutional framework through which material streams are defined, regulated, and reported. Waste management is not governed by a single authority, but rather by an overlapping system of federal statutes, state agencies, and regional enforcement bodies, each organized around distinct environmental risks and material pathways. These institutional distinctions are central to the present research because they shape both the observable infrastructure network and the reporting systems from which empirical data are drawn. Indeed, national case studies have shown that

treatment pathways are shaped by regulatory frameworks, market structures, and infrastructure capacity constraints (Cayumil et al., 2021).

At the federal level, waste regulation is anchored in the Resource Conservation and Recovery Act (RCRA) of 1976, which established the foundational “cradle-to-grave” framework governing the generation, transport, treatment, storage, and disposal of both solid and hazardous waste. Complementary statutes, including the Clean Air Act, Clean Water Act, and Toxic Substances Control Act, regulate emissions, discharges, and toxic constituents associated with waste facilities and downstream environmental releases. Together, these laws provide a national scaffold, but implementation remains highly decentralized, leaving substantial regulatory discretion to states, resulting in considerable state to state variability.

California has developed one of the most comprehensive state-level waste governance systems in the United States under the umbrella of the California Environmental Protection Agency (CalEPA). Within this structure, regulatory responsibility is distributed primarily across three institutions according to material stream and environmental medium (CalEPA, 2026).

For non-hazardous municipal solid waste, recycling, and organics diversion, the principal agency is the California Department of Resources Recycling and Recovery (CalRecycle). CalRecycle oversees landfill permitting and closure, diversion programs, beverage container recycling, organics mandates, and extended producer responsibility programs. Its jurisdiction spans the main material streams examined in this dissertation (with the exception of e-waste), including municipal solid waste, recyclable commodities, compostables, construction and demolition materials, and producer-managed packaging streams.

Hazardous, toxic, and universal waste streams fall under the authority of the Department of Toxic Substances Control (DTSC). DTSC governs the generation, transport, treatment, and final disposition of hazardous materials, including industrial chemicals, solvents, batteries, and certain electronic wastes. California's hazardous waste framework is notably broader than the federal baseline, applying state-specific criteria that capture some materials that are not classified as hazardous under federal law.

A third major regulatory domain concerns environmental releases to land and water. The State Water Resources Control Board and the nine Regional Water Quality Control Boards regulate landfill liner systems, leachate management, groundwater monitoring, compost runoff, and other waste discharges through Waste Discharge Requirements (WDRs). This jurisdictional layer is particularly important for disposal facilities, where waste regulation intersects directly with groundwater and surface water protection.

California's contemporary waste governance system is best understood as the result of successive legislative generations, each expanding the scope of regulation from disposal control to diversion, climate mitigation, and ultimately lifecycle material stewardship. While the federal framework established foundational environmental protections, California has progressively extended that framework to address system-wide material recovery, emissions reduction, and producer accountability. The evolution of this framework can be understood in three broad legislative phases: first-generation diversion policy, second-generation systems and climate integration, and third-generation circular economy and extended producer responsibility (CalRecycle, 2026b).

The foundational statute is the California Integrated Waste Management Act of 1989 (AB 939), which fundamentally reoriented state waste policy away from almost complete reliance on

landfills and toward source reduction, recycling, and composting. Prior to AB 939, California lacked a cohesive statewide strategy for managing rapidly increasing municipal solid waste volumes. The Act imposed mandatory diversion targets requiring local jurisdictions to achieve 25% diversion by 1995 and 50% by 2000, thereby assigning cities and counties direct responsibility for waste reduction planning and implementation (CalRecycle, 2026c).

AB 939 institutionalized the waste hierarchy of source reduction, reuse, recycling, composting, and disposal as a final resort, while simultaneously establishing the permitting, planning, and reporting infrastructure that underpins California’s modern waste system. Requirements for Integrated Waste Management Plans (IWMPs), minimum disposal capacity planning horizons, and state funding mechanisms through landfill dumping (“tipping”) fees created the administrative structure later inherited by CalRecycle.

This first generation also includes state environmental statutes that govern specialized waste pathways and environmental media. The Porter-Cologne Water Quality Control Act (1969) established California’s broad authority over discharges affecting both groundwater and surface water, including landfill leachate and compost runoff (CSWRCB, 2026). Similarly, the Hazardous Waste Control Law (1972) created a state-level cradle-to-grave hazardous waste framework that predates and exceeds the federal Resource Conservation and Recovery Act (RCRA).

Together, these statutes define the foundational regulatory architecture governing disposal, diversion, hazardous materials, and waste-related environmental releases. The second legislative generation expanded California’s waste policy beyond municipal diversion targets toward statewide systems management and explicit climate integration.

A major milestone in this phase is AB 341 (2011), which established a statewide goal of 75% diversion and expanded mandatory recycling requirements into the commercial sector. This legislation marked a transition from local diversion compliance toward statewide systems optimization and infrastructure development.

This expansion continued with AB 1826, which introduced phased requirements for commercial organics recycling. By explicitly targeting food and green waste streams, AB 1826 recognized the disproportionate role of organics in landfill methane generation and the need for dedicated recovery infrastructure such as composting and anaerobic digestion (CalRecycle, 2026d). The culmination of this phase is SB 1383 (2016), which integrated waste regulation with California's climate policy objectives. SB 1383 mandated a 75% reduction in organic waste disposal relative to 2014 levels, along with enforceable edible food recovery requirements (CalRecycle, 2026e). This legislation was a decisive shift from programmatic encouragement toward prescriptive and enforceable methane mitigation policy.

An equally significant institutional development was AB 901 (2015), which created the statutory basis for the Recycling and Disposal Reporting System (RDRS). By requiring quarterly reporting of types, quantities, and destinations of materials, AB 901 transformed California's waste governance into a data-intensive reporting system capable of statewide material flow reconstruction (CalRecycle, 2019). This legislation directly underpins the empirical framework used in this dissertation.

The third generation of California waste legislation extends regulation upstream, from management of generated waste to the design, composition, and lifecycle responsibility of products themselves. This phase is defined by the rapid expansion of extended producer responsibility (EPR) and circular economy legislation.

Most prominently, SB 54 established one of the most comprehensive packaging EPR systems in the United States, requiring all packaging to be recyclable or compostable by 2032, mandating a 25% reduction in plastic packaging, and shifting responsibility for mitigation and recovery costs to producers (CalRecycle, 2026f).

Related statutes include:

- SB 1215, which expanded producer responsibility for electronic waste (CalRecycle, 2026g);
- AB 2440, which introduced battery stewardship requirements (CalRecycle, 2026h);
- SB 270, which regulated single-use plastic bags (CalRecycle, 2026i); and
- SB 343, which strengthened truth-in-labeling requirements for recyclability claims (CalRecycle, 2026j)

This legislative phase reflects a broader conceptual transition from downstream waste management toward closed-loop material governance, in which policy explicitly targets both the generation and fate of material streams. California's framework increasingly incorporates an integrated systems perspective, linking product design, recovery infrastructure, secondary material markets, and climate outcomes into a unified regulatory logic.

2.2 Data Components of California's Regulatory Framework

Two complementary databases are central to this research: the Solid Waste Information System (SWIS) and the Recycling and Disposal Reporting System (RDRS). Together, these databases provide a regulatory inventory of facilities and their permitted activities and directional reporting of material movements between entities, respectively. While each database has some limitations, their integration enables reconstructing California's post-collection waste network at statewide

scale, providing empirical origins and destinations of waste movements between facilities, as well as some flows of recyclables into export and brokerage markets. The following subsections describe each dataset in detail, including structure, scope, and limitations relevant to system-level modeling.

2.2.1 Solid Waste Information System (SWIS)

SWIS is a relational database that contains detailed information on permitted waste processing and disposal activities across California (CalRecycle, 2026k). It includes data on 5,655 sites (both active and closed), such as location (address, latitude, longitude), contact information of owners and/or operators, permitted throughput quantities, and other regulatory details such as relevant reporting and enforcing organizations, Boolean variables indicating various conditions of compliance, names of internal case managers/inspectors, and actual or forecasted closure dates. The data are organized into five tables (Sites, Activities, Owners, Operators, Waste), all linked by a unique identifier known as the ‘SWIS number’ assigned to each site.

Each site in SWIS may have multiple permitted activities, including Composting, Disposal, Engineered Municipal Solid Waste (EMSW) Conversion, In-Vessel Digestion, Transfer/Processing, and Transformation. These activities are characterized by permitted throughput and capacity. This information is critical for understanding site operations. Activities are classified by operational and regulatory status, similar to site-level categories. I verified that activities marked as ‘Active’ (operational status) are associated with active sites. Activities at active sites were compiled into a lookup table that includes available quantitative data on permitted throughput (typically mass per time). These appear consistent with the site type categorization, which differentiates between disposal and non-disposal activities, with some sites handling both types of activities. Examination of reported materials passing in, out, or through a

given facility in a given time period reveals two notable issues: 1) reports do not necessarily indicate all permissible activities (i.e. processing construction debris inerts (C&D) is permitted but the site does not appear to do so) and 2) that permits are typically set at levels far above the actual processing throughput. Thus, permit data alone is a poor proxy for reported operational data.

Additional tables in SWIS include Owner, Operator, and Waste tables. The Owner and Operator tables can be useful for identifying points of contact at facilities or for finding sites associated with specific companies. However, the ongoing flux in this information makes the currency of information uncertain, and it may not reflect the actual state for specific time periods. I experienced this when I contacted some facilities – many phone numbers were disconnected and some contacts had changed. SWIS waste data tables describe the types of waste permitted at each site. For instance, a monofill operation (i.e. a specialized, engineered landfill designed to accept only one specific type of non-hazardous waste material) might only accept specific materials like ash or tires, whereas a Class I landfill may accept nearly all types of waste, including hazardous materials. This table is generally less useful for broader analyses but it can confirm constraints on some landfills or the lack thereof.

The primary limitation of SWIS is that it only records sites that handle mixed municipal solid waste. It does not include downstream users or processors of recyclables or other diverted materials. It also does not describe some permit details such as apparently omitted activity-specific permit data and under what conditions operators report under combined permits. Furthermore, at the end of 2023, while approximately 1200 SWIS sites were active, only half of those could be joined with high fidelity to RDRS entities. This gap does not reflect a problem

with SWIS so much as it is a symptom of RDRS that is still a maturing database that is not fully cross-referenced.

2.2.2 Recycling and Disposal Reporting System (RDRS)

RDRS was created by Assembly Bill 901 (AB 901) in 2015. AB 901 required businesses to “report directly to CalRecycle on a quarterly basis on types, quantities, and destinations of materials... disposed of, sold, or transferred inside or outside of the state”. The implementation of AB 901 (and the associated Public Resources Code §41821.5) necessitated a new reporting system and database which came online in 2019 (CalRecycle, 2026l). RDRS consists of reports compiled from facilities that handle more than 100 tons per quarter, as required by CCR Title 14 Division 7 Article 9.25. These facilities, which are registered as "entities" in the RDRS database, include landfills, transfer stations, recyclers, composters, processors, brokers, and haulers. This system, which was launched in mid-2019, provides a comprehensive view of California’s waste infrastructure. RDRS superseded DRS which did not include a full range of entities or materials. Data in RDRS are divided into public and privileged, on the basis of material and facility type.

Public data are accessible via several key reports. CalRecycle provides lists of reporting sites/entities with names, locations, status, and activities, referenced by RDRS IDs. There are 2,804 unique RDRS IDs (as of April, 2024), although some facilities have multiple records due to multiple activities. Table 2.1 displays the frequency of each reporting entity activity label in RDRS.

Table 2.1 Reporting Entity Type and Frequency

Reporting Entity Label	Count
Broker/Transporter	469
Contract Hauler	87
Engineered Municipal Solid Waste Conversion	5
Food Waste Self Hauler	13
Landfill	167
Other Disposal	71
Recycler/Composter	1222
Transfer/Processor	747
Transformation	23

Also essential for this work is Jurisdiction Disposal Totals, which provides the tonnage of solid waste disposed of in landfills, organized by jurisdiction of origin (primarily cities, counties, and public waste management authorities). It also indicates whether and how much waste was passed through transfer stations. Even though this report only covers municipal solid waste, it has strong modeling potential as it describes complete waste flows from collection to disposal.

The remaining public reports (1-5, and 7-9) are highly aggregated (e.g., statewide, by county, jurisdiction, or facility) or pertain to aspects of waste management less relevant to my specific research questions.

The REMADE project, which motivated this work, obtained RDRS privileged data via a no-disclosure agreement (NDA) with CalRecycle. We received two main tables: “Entities” and “Outflows.” The entities table is identical to RDRS Report 0. It lists facilities, their locations, and contact information for all registered entities, including recyclers, transfer stations,

processors, and brokers. The outflows table details material flows between entities from 2019 Q3 to 2023 Q1. It records origin (focal) and destination facility identifiers, and includes data on various material streams and categories, with over 1,000 combinations that can be mapped to more general commodity categories. Quantities are recorded in tons.

There are several issues with the RDRS data. First, as could be expected for a document resulting from the reports of thousands of different entities, not all of the data have been fully checked for fidelity upon entry. It appears that some facilities were assigned new RD identifiers and in general, RD numbers are not unique – it takes the merging of three text fields into a bulky “triumvirate” key to be unique. And this must be unified across RDRS, SWIS, and DRS, a painstaking task to ensure the high fidelity of connections.

Second, the address provided does not always correspond to the physical location of where waste or recycling activities are taking place, as different entities from the same organization can share the same address (e.g. a city sanitation department or other office/headquarters). This is a prevalent issue with brokers, as many of the addresses correspond to commercial/retail or even residential areas rather than a commercial or industrial site.

Third, the interpretation of certain fields (e.g. host assigned tons), administrative tags (e.g. exempt, provisional), and systematically missing data is complicated by the fact that this database is relatively new. This means that CalRecycle is still working to establish fair rulemaking that is useful to regulators but not so onerous to collect as to prevent reporting. Trying to satisfy the numerous stakeholders who would be generally opposed to more reporting work, if not more government oversight, is obviously a very difficult task (CalRecycle, 2013). Through several conversations, CalRecycle employees indicated that refinements in rulemaking and implementation as well as engaging more sectors (e.g. dairy farms) have taken priority as

their regulatory mandate continues to expand. However, there has been progress in closing change-control-related gaps such as repairing some missing or erroneous identifiers. The proximate problems with RDRS data fidelity are not insurmountable – uniqueness can be forced, addresses can be geocoded - but uncertainty in the dynamics of the database implementation hampers the creation of system boundaries.

2.2.3 Data Limitations

Some key data limitations need to be noted for this work: waste entry and waste exit, and inadequacy of regulatory facility type labels.

The main data reporting issue that must be understood and that this research overcomes is related to the lack of material balance for a given facility. There should be no expectation of net zero material balance. Landfills, by definition, accumulate material, but there is not a specific accounting structure tracking waste. The reporting entity reports outflows and identifies the receiving facilities. However, there is no guarantee that the receiving facility will report equivalent data. Even more problematic is that once a target recyclable material is recovered and concentrated into a secondary commodity (e.g. a bale of pure aluminum cans), it may no longer be considered waste so it no longer needs to be reported in RDRS. This accounting gap of disappearing and reappearing of waste could be described as “leakage” and “seepage” of data.

Furthermore, waste entering the system from municipal collection or self-haul is systematically unreported unless it originates from a registered entity. Therefore, some inflows to facilities, like transfer stations, are not explicitly accounted for.

Likewise, not all waste exits are directly reflected in RDRS. End use material streams capture disposal and recycling activities reported without a specific destination. Exports are

similarly identified but with generalized destination labels like “Out of CA” or “Out of USA.” Ambiguities exist when inflows are reported without corresponding outflows, suggesting incomplete reporting, though these are perhaps an intentional artifact of the negotiation with operators, as discussed earlier.

The combined impact of these various reporting conventions confounds any attempt to apply a traditional mass balance of “in minus out plus generation minus consumption equals accumulation” to the empirical data. Nevertheless, RDRS gives an unprecedented window into the waste management organization in terms of scale and breadth, and CalRecycle’s data disclosure to the UCI REMADE team provided a valuable research opportunity.

Another problem with RDRS is that the stated facility types are not well-defined for distinguishing operational profiles. Some facilities have incredibly diverse activities, acting as a one-stop-shop for transfer, recovery, composting, and/or landfilling. This becomes problematic because, for convenience of reporting, these activities are rarely separated into their components. This is likely because intermediate weights of subprocess flows (e.g. residuals from a MRF line) are simply not collected prior to being aggregated with transload/MSW material. These kinds of facilities typically have only inbound and outbound scale-houses (a scale house is a small building next to an industrial scale that serves as a control office for employees to weigh trucks or track inventory) and they are not about to load trucks and drive around the block to weigh them. The other side of that same coin is that a landfill that combines its reporting with an on-site MRF appears to divert recyclable materials. However, there is no way to separate the residuals of their MRF operation from the overall disposal number or to know the relative size of the operations, confounding attempts to determine recovery efficiency.

While these datasets define the institutional structure and reported flows of California's MSW system, they do not encode the operational meaning of material movements or facility activities. To interpret these data, we first need to define the characteristics of the materials being managed and the roles of different facilities within the system.

2.3 Material Classes and Processing Pathways

Material properties fundamentally govern processing requirements, recovery potential, and routing within the waste management system. Differences in density, contamination tolerance, chemical composition, and market value dictate both the technologies needed for separation and the economic feasibility of recovery. Recyclable materials in municipal waste streams originate from a diverse set of sources and undergo a sequence of sorting, processing, and remanufacturing steps that ultimately determine their quality, grade, and end-use applications.

According to CalRecycle, the California municipal waste stream is primarily composed of organics, paper, plastics, and inerts, with also a substantial percentage of metals, glass, and miscellaneous materials.

Organic materials include food waste, landscape trimmings, lumber and sludges. These materials are central to efforts to mitigate methane emissions by processing them through composting and anaerobic digestion rather than landfilling. They represent ~28.4% by weight of the materials that entered California landfills in 2021 (CalRecycle, 2024).

Paper and cardboard is the second largest category of materials sent to landfills in California (~15.5% by weight in 2021, or ~6.19 million tons). This material stream is generated across residential sectors, commercial and institutional establishments, industrial facilities, and supplemental collection systems such as recycling drop-off centers or public bins. These streams

are typically separated into material grades, with higher-quality fibers such as office paper and stationery distinguished from lower-grade materials such as old corrugated cardboard (OCC). At recycling facilities, recovered paper is directed to end users such as paper mills and packaging manufacturers, where it undergoes pulping in systems analogous to industrial blenders.

Plastics are predominantly sourced from curbside recycling programs, dedicated recycling centers, deposit-return systems for beverage containers, and specialized collection programs such as plastic bag take-back. In California, plastics comprise ~13.7 percent of landfilled material by weight (~5.45 million tons) in 2021 (CalRecycle, 2024). The containers and packaging category represents the largest share of plastic waste, including PET bottles, HDPE containers, films, and wraps. Among these, polyethylene terephthalate (PET) is a major recyclable resin, particularly in post-consumer packaging. Mechanical recycling is the most common pathway for PET, involving intensive sorting to achieve resin purity, followed by size reduction, washing, and reprocessing into pellets or flakes for reuse in manufacturing.

Construction & demolition (C&D) and inert debris waste includes concrete, dirt, and miscellaneous construction debris. The mixed materials are often referred to as “inerts” because materials such as concrete, asphalt, wood, drywall, and metals dominate. They represented ~12.0% of materials landfilled in California in 2021 (CalRecycle, 2024).

Metals, by comparison, made up only ~4.9% of flow of materials sent to landfills, partly because municipal solid waste does not include metals from junked vehicles or discarded appliances, which are the main source of metal recycling in California. It is customary to distinguish between **ferrous metals** (primarily iron and steel, ~1.96 million tons in 2021), and **non-ferrous metals** (primarily aluminum, copper, and lead). Ferrous metals are recovered from a wide array of sources, including construction and demolition (C&D) debris. A defining

characteristic of ferrous metals is that they can be recycled repeatedly without degradation of material properties, although preprocessing steps such as de-tinning may be required depending on the input stream (CalRecycle, 2024). Non-ferrous metals are typically more heterogeneous and higher value than ferrous metals, necessitating more precise sorting and separation technologies to meet end-user specifications.

Glass is next, with ~2.3% by weight of materials sent to California landfills in 2021 (CalRecycle, 2024). Collected glass is first sorted—often by color and contamination level—before being crushed and cleaned to produce cullet, a uniform granular feedstock. This cullet is then supplied to glass manufacturers, where it is melted in high-temperature furnaces and formed into new products. Like ferrous metals, it can be repeatedly recycled without loss of quality. Recycled glass is a critical input in closed-loop container recycling systems.

Finally, **miscellaneous materials** (e.g., textiles footwear, linens, aseptic packaging, personal care items, organic grit, or treated wood that cannot be composted) made up ~16.9% by weight of the materials sent to California landfills in 2021 (CalRecycle, 2024).

2.4 Waste Management Infrastructure and Facility Types

Waste management in California is a complex web of individual waste facilities of various types that represent the structural nodes of the waste management system. While regulatory classifications define permitted activities, they only partially reflect the operational roles facilities play in processing different material streams. The post-collection waste management system consists of a network of interdependent facilities that each serve specialized roles in sorting, recovering, transferring, and disposing of waste. These facilities operate as components of an integrated system designed to maximize resource recovery for recycling and minimize environmental harm. Their design, business models, and material handling capabilities reflect

both regulatory frameworks and market incentives, and each interacts differently with mixed versus source-separated waste streams.

Facility types include: 1) Transfer Stations and Processors; 2) Material Recovery Facilities; 3) Organics Material Management Facilities; 4) Construction Debris Inert Processing Facilities; 5) Landfills; and 6) Broker Facilities. These will be described individually in the following subsections.

2.4.1 Transfer Stations and Processors

Transfer Stations typically receive materials directly from collection routes (“first touch” materials) whereas Processors may receive materials that have been previously source-separated (“second touch” materials or commercial contracts). Therefore, the designations of Transfer versus Processor facilities cannot be determined without analyzing reported material flows.

Transfer stations are foundational to the logistics of waste management. Typically located closest to waste generation points than landfills or other large processing centers, these facilities consolidate materials collected from residential, commercial, and industrial sources. Their primary function is to improve the efficiency of hauling by transferring waste from smaller collection trucks to long-haul vehicles that transport it to more distant disposal, recycling, or processing sites. Transfer stations typically charge fees on a per-ton basis to individual haulers. This is commonly referred to as a “tip fee” in the industry. Many modern transfer stations incorporate limited sorting functions or contamination checks, thus immediately diverting bulky, hazardous, and/or recyclable materials away from landfills. Their role is essential, serving as the connection between primary collection routes and downstream processing or disposal.

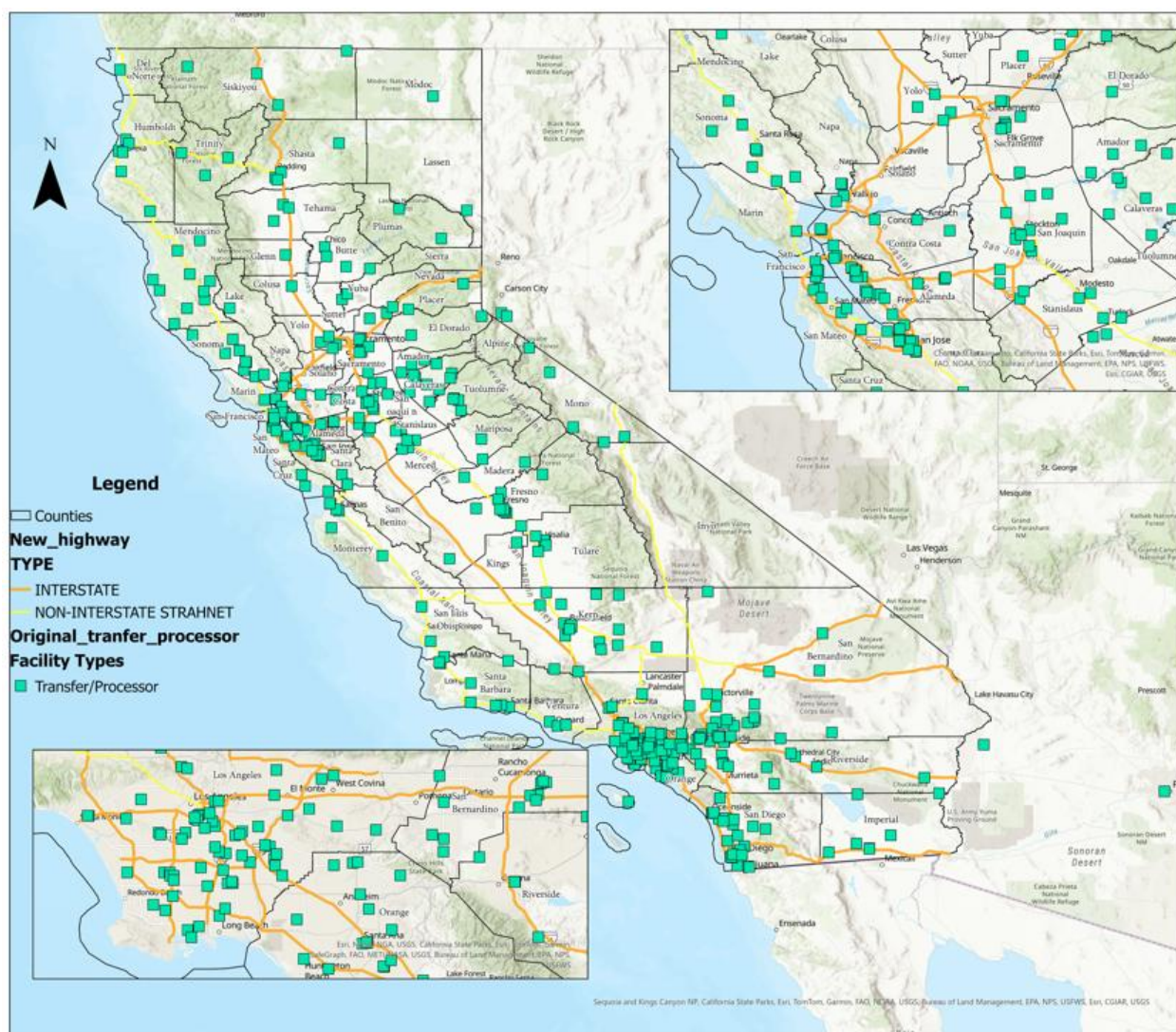


Figure 2.1 Map of Transfer Stations and Processors

Figure 2.1 shows the locations of transfer stations and processors (RDRS entities reporting in 2021). As expected, they are concentrated primarily within populated areas throughout the state. The designation of Transfer/Processor by CalRecycle’s RDRS database is unfortunate as the difference in naming denotes a regulatory nominal 40% cap on contamination flowing into “processors” to be defined as such. The streams of commercial, source-separated, or pre-processed recyclables are implicitly higher value than those coming directly from collection

trucks. Not only is there more potential commodity, but less effort to purify them due to less contamination. This is beneficial for consistency of supply and economies of labor, scale, and reliability.

The logistical importance of transfer stations has been widely recognized in optimization and routing studies, which show that consolidation reduces the total vehicle miles traveled but can shift emissions spatially depending on downstream distances (Andeobu et al., 2022). Empirical analyses emphasize that collection and transport often dominate operational budgets, particularly in geographically dispersed systems (Cayumil et al., 2021). These dynamics underscore the importance of accurately characterizing the contribution of transfer stations in system-level emissions modeling.

2.4.2 Material Recovery Facilities

Material Recovery Facilities (MRFs) are the backbone of diversion efforts and are the most complex nodes in the recovery network. Their business models center on capturing value from commingled recyclables by sorting and processing them into marketable commodities (Karter, 1993). MRFs receive single-stream recycling materials collected from residential curbside programs and source separated material from commercial entities. These materials are sent through a combination of mechanical and manual separation processes. Optical sorters, eddy current separators, magnets, screens, and air classifiers, as well as manual sorting systems, are commonly used to sort materials based on physical properties such as size, shape, density, magnetic susceptibility, and spectral reflectance. Ferrous metals are extracted using magnetic separators; eddy current systems repel nonferrous metals like aluminum from the stream; optical sorters distinguish plastic resins by light reflection; and screens isolate fiber products by planar geometry. The facilities then produce sorted bales of materials for sale to domestic or overseas

processors. Their revenue is tied both to tipping fees and the commodity value of outputs, making certain materials highly sensitive to market prices and contamination levels.



Figure 2.2 Map of MRF locations

Life-cycle assessments consistently identify recycling of metals and fibers as major contributors to net environmental benefits, while contamination rates and residual disposal materially affect outcomes (Andreasi Bassi et al., 2017). A comparative study in the European

Union showed substantial variation in recovery efficiency and residual rates across facilities (Ríos and Picazo-Tadeo, 2021). These findings highlight the need to differentiate MRFs by operational characteristics rather than administrative labels.

However, MRFs are not separately identified as a category and could be considered either Processors or Recyclers under the baseline RDRS categorizations, making analysis more complicated. CalRecycle generously disclosed an in-process survey of MRFs throughout California. This information, in tandem with RDRS and SWIS, were leveraged to produce a list of MRFs under fairly strict definitions in terms of material streams and reporting consistent with MRF operations. Even CalRecycle, through their surveying, reached out to facilities which turned out not to be MRFs by their definition or not active. Figure 2.2 shows the locations of the MRFs identified.

2.4.3 Organics Material Management Facilities

Organics material management facilities include composting and anaerobic digestion (AD), and serve as the primary destination for organic materials diverted from the waste stream. Traditional composting relies on aerobic decomposition of food scraps, yard trimmings, and compostable paper to produce soil amendments. These facilities vary in scale from small open-windrow sites to large in-vessel systems capable of handling contaminated commercial food waste streams. AD facilities are specialized subset, use anaerobic microbial activity to convert organics into biogas and digestate. The biogas is often captured for energy use, qualifying these systems for renewable energy credits, while the digestate may be land-applied as a soil amendment. The viability of composting facilities depends on their ability to receive clean feedstock, making them more compatible with source-separated organics programs than with commingled MSW. Their role in the system is increasingly critical due to organic landfill bans and methane

reduction goals, with revenue often derived from tipping fees, compost sales, and energy credits in the case of AD facilities (Bux et al., 2023; Manea et al., 2024; Onwosi et al., 2017; Wang et al., 2020).

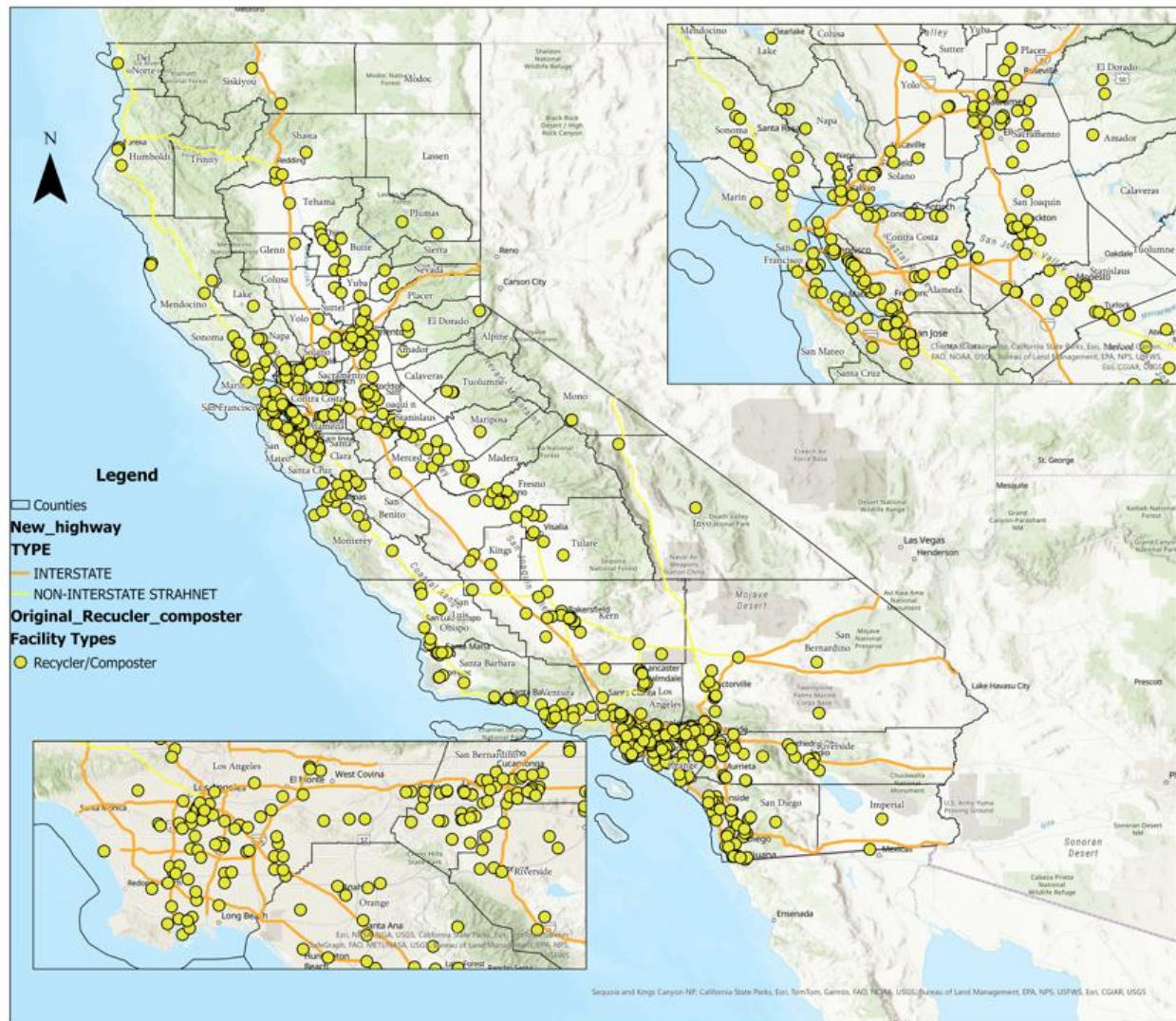


Figure 2.3 Map of Recyclers and Composters

Figure 2.3 shows the locations of recyclers and composters in California. It is difficult to draw conclusions about their spatial distribution because they appear to be simply everywhere, clustered in and around cities, highways, waterways. Notably, the San Francisco Bay inset shows

the entire south bay industrial area from San Jose to Oakland positively blanketed. Similarly, the South-Central Los Angeles, Riverside, and San Bernardino regions in Southern California are covered with multiple facilities. However, it is also apparent that, in comparison to transfer stations, there are many facilities located on the outskirts of populated areas, where there are larger land spaces required for large scale composting. Composting requires a significant amount of land space. Again, it is problematic for interpretation that the recycling and composting facilities cannot be separated without further analysis of their reporting.

Organics diversion is increasingly central to methane mitigation strategies worldwide (Chen et al., 2020), National and regional analyses demonstrate that composting and anaerobic digestion performance depends strongly on feedstock contamination, energy substitution context, and system boundaries (Andreasi Bassi et al., 2017). These factors must be explicitly represented when evaluating diversion impacts.

2.4.4 Construction Debris and Inert Processing Facilities

Construction & demolition (C&D) and inert debris recovery facilities target waste loads from construction sites and renovation projects. The mixed materials are often referred to as “inerts” because material types like concrete, asphalt, wood, drywall, and metals dominate. Processing involves crushing, screening, and manual sorting, to recover aggregates for reuse and extracting ferrous metals. Ferrous materials are typically sold into the scrap market, while crushed clean concrete may be reused as base material. Business models often depend on lower tipping fees than landfills and sales of recovered materials. While less integrated with residential waste systems, these facilities intersect with transfer stations when mixed C&D waste is accepted and rerouted. They are critical for meeting diversion mandates for construction waste.

The waste management literature highlights the importance of crushing, screening, and recovering metals from C&D debris for achieving diversion targets, with significant variability in aggregate reuse markets and reporting practices (Andeobu et al., 2022). This variability reinforces the need for empirical classification based on observed material flows.

2.4.5 Landfills

Landfills are the final repository for non-diverted waste and remain necessary even in highly optimized systems. Modern landfills are engineered for mitigation of environmental containment, with leachate collection and methane gas capture systems.

Landfill classification in California is structured as a coupled system in which waste types are first categorized by their potential to impair water quality, and disposal facilities are then classified according to the level of containment required to safely manage those wastes. Under CCR Title 27, designated wastes with elevated leachate risk must be placed in Class II landfills featuring more stringent geologic or engineered barriers, while nonhazardous municipal solid waste is directed to Class III landfills with comparatively simpler liner and containment systems (California Office of Administrative Law, 2026). Maximal containment is achieved in Class I facilities, designed for hazardous waste disposal. Inert wastes, characterized by negligible decomposition and leachate generation, are subject to minimal regulatory controls. This framework therefore functions as a risk-matching scheme, where increasing waste hazard is met with increasing containment performance, and facility classification reflects environmental protection requirements rather than operational or processing characteristics. The SWIS database stores information on landfill class.

The observed distribution of landfill classifications is strongly skewed toward lower-risk disposal categories (CalRecycle, 2026k). Class III landfills comprise the majority of units (115

facilities), reflecting the predominance of nonhazardous municipal solid waste disposal under standard containment requirements. Inert waste units also represent a substantial share (44 facilities), consistent with the presence of facilities or cells dedicated to materials with negligible decomposition potential. By comparison, Class II landfill units are relatively limited (25

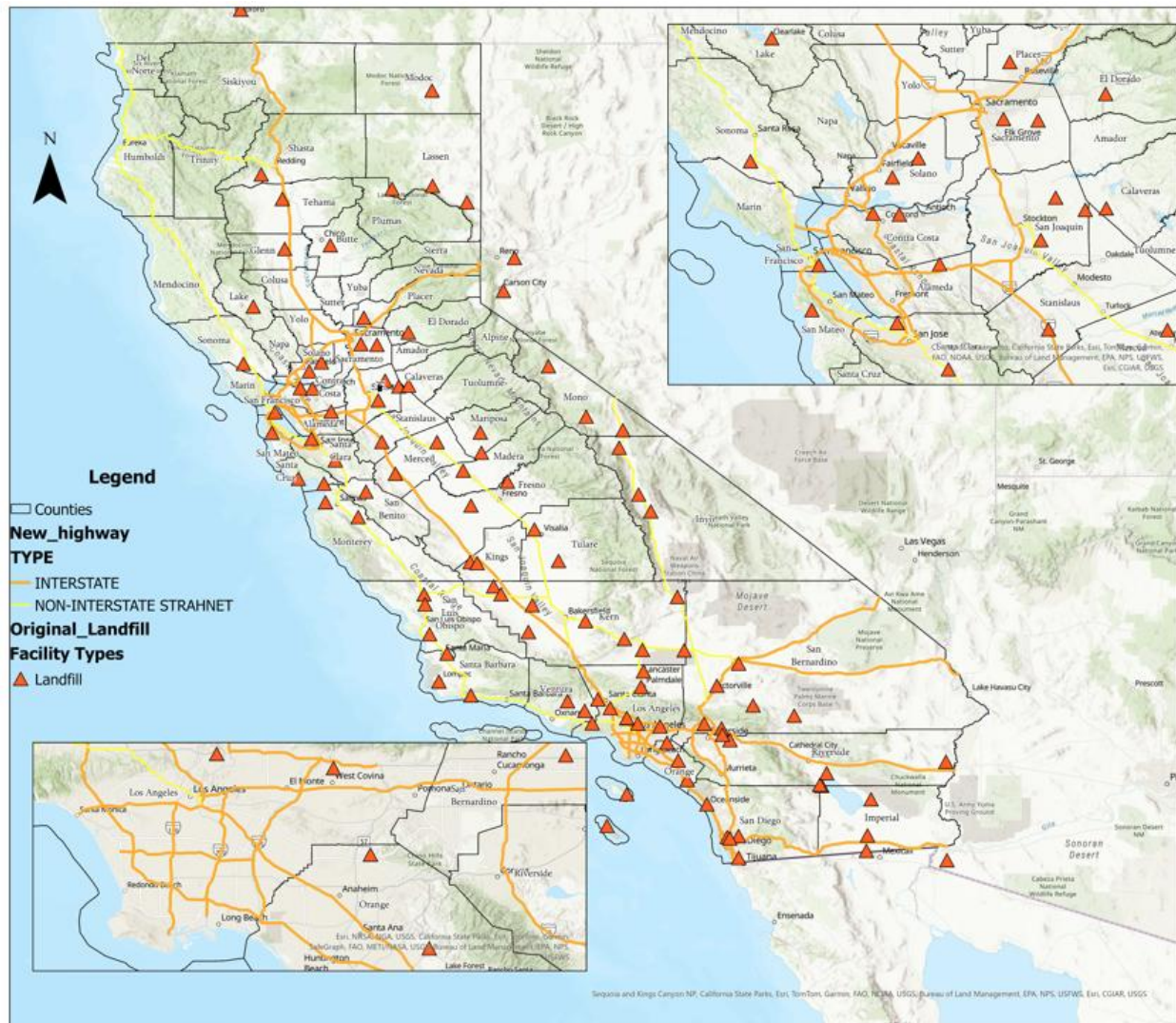


Figure 2.4 Map of Landfills

facilities), indicating a smaller subset of facilities managing designated waste requiring enhanced containment. As expected, Class I landfill units are rare (4 facilities), given the highly

specialized nature of hazardous waste disposal. It is important to note that these classifications do not necessarily apply to entire facilities; rather, they are often assigned to individual waste management units or landfill cells within a larger facility, so a single facility may contain multiple cells with different classifications.

Figure 2.4 displays the locations of landfills (RDRS entities reporting in 2021). Facilities are more dispersed and tend to be away from highly populated areas. Some out-of-state landfills are visible near the state border, as they accept waste from California jurisdictions.

Many facilities now operate as part of larger vertically integrated systems managed by waste companies that also own or operate MRFs and transfer stations. Their revenue comes from tipping fees, and some generate additional income by selling landfill gas or accepting special waste. Landfills often receive residuals from other facilities—such as rejects from MRFs or contaminated organics—and must manage material flow that cannot be economically, technologically, or safely diverted. They serve as a crucial pressure valve in the integrated system, especially when markets for recyclables contract or contamination renders recovery unviable as well as a last opportunity for diversion. While not typically involved in sorting, some sites report outflows of easy to identify recoverable material such as mattresses or scrap metal.

Landfill methane remains one of the largest waste-sector greenhouse gas sources globally (Chen et al., 2020). Cross-country analyses show that diversion gains can be offset by persistent landfill reliance and long-term gas generation (Tisserant et al., 2017).

2.4.6 Brokers

Waste brokers function as intermediaries within the materials market, rather than processors themselves. Their role is to coordinate the sale and shipment of sorted commodities (recyclables,

recovered metals, fiber, plastics) to or from MRFs and other facilities to end-markets, both domestic and international. Brokers often work on commission or margin, profiting from arbitrage between facility output pricing and market demand. In an integrated system, they provide the linkage that determines whether recovery efforts translate into actual diversion, facilitating transactions, improving quality of commodities, and connecting markets.

When reviewing the entity database for geocodes, it became apparent that many of the addresses for brokers did not correspond to a hauling yard, so they are not mapped here. Many are residential locations, small commercial offices, or even PO boxes. Careful examination of broker geocodes indicates that approximately half of the sites are not locations where any waste would actually travel to/from. Thus, reported inflows and outflows to many broker facilities cannot be included in analysis of actual material movements, because the transfer only indicates a change in ownership of the materials.

Unfortunately, brokers and other middle operators do not want their network known. Throughout all reporting for the state in 2021, only 29 brokers filed their own report as an entity in CalRecycle (and again, about half are artificial locations). There are broker “destinations” in 20 other states, as well as entities apparently operating out of Canada, Italy, Switzerland, the Netherlands, Pakistan, India, Indonesia, Hong Kong, and Taiwan.

2.5 Empirical Material-Facility Interactions

With the material characteristics and nominal facility roles established, the RDRS data can be interpreted as empirical observations of how materials are routed through California’s waste and recycling infrastructure. These reported flows reveal both expected patterns consistent with facility design and systematic deviations that reflect operational complexity, reporting conventions, and the multi-stage nature of material processing.

Records of material movement in RDRS are best understood as quantities of designated material streams transferred from an origin activity to a destination activity. Rather than representing a strict mass balance at the facility level, these records capture reported transfers between administrative entities. Consequently, the data provides a network representation of transport and handling activity rather than a fully traceable chain of custody for individual waste shipments.

The reported movement of materials across the California waste and recycling system reveals substantial heterogeneity in the relationship between origin activity, destination activity, and declared material stream purpose. Rather than following a simple one-to-one correspondence between facility type and operational role, the data in Table 2.2 indicate a highly interconnected system in which facilities frequently participate in multiple functional pathways.

For solid waste for disposal, the dominant pathway is movement from transfer/processors to landfills, totaling approximately 24.0 million tons annually, confirming the central role of transfer infrastructure in consolidating and routing disposal-bound waste. This is consistent with the expected function of transfer stations as intermediary nodes that receive collection vehicles and dispatch consolidated loads to final disposal facilities. However, substantial quantities also

Table 2.2 Material Streams between Activities (2021 short tons)

Material Stream	Origin Activity	Destination Activity				
		Landfill	Transfer/Processor	Recycler/Composter	Other	Unspecified/Broker/Transporter
Solid Waste for Disposal	Landfill	2,244	12,899			
	Transfer/Processor	24,042,192	298,298		364,451	
	Recycler/Composter	1,086,718	106,168		39	
	Broker/Transporter	271,974	3,680			

	Other	5,128	2,940			
Beneficial Reuse	Landfill	2,129	5,218			
	Transfer/Processor	1,231,754	288,206			
	Recycler/Composter	519,019	55,495			
	Broker/Transporter		6380			
	Other	101,078				
Recycling/Composting	Landfill	3,207	157,083	338,162	3,065	6,592
	Transfer/Processor	664,565	2,113,319	3,411,702	40,661	398,480
	Recycler/Composter	256,822	400,574	2,004,946	70,540	521,698
	Broker/Transporter	25,569	221,612	698,045	6,144	439,109
	Other	1,327	819	13,292		20,730
Brokering/Transporting	Landfill		23		368	78,324
	Transfer/Processor		4,607		82	790,925
	Recycler/Composter		15,806	37		1,461,339
	Broker/Transporter		298			373,792
Designated Waste for Disposal	Transfer/Processor	33,059	4,693			
	Other		329			
End Use	Landfill					172,952
	Transfer/Processor					1,779,248
	Recycler/Composter					10,078,281
	Broker/Transporter					1,019,827
	Other					72,107

originate from recycler/composter activities (approximately 1.09 million tons) and broker/transporter activities (approximately 0.27 million tons), indicating that residuals and unrecovered fractions from diversion-oriented facilities ultimately re-enter the disposal stream. Figure 2.5 indicates the distribution of outflows for Transfer/Processor activities.

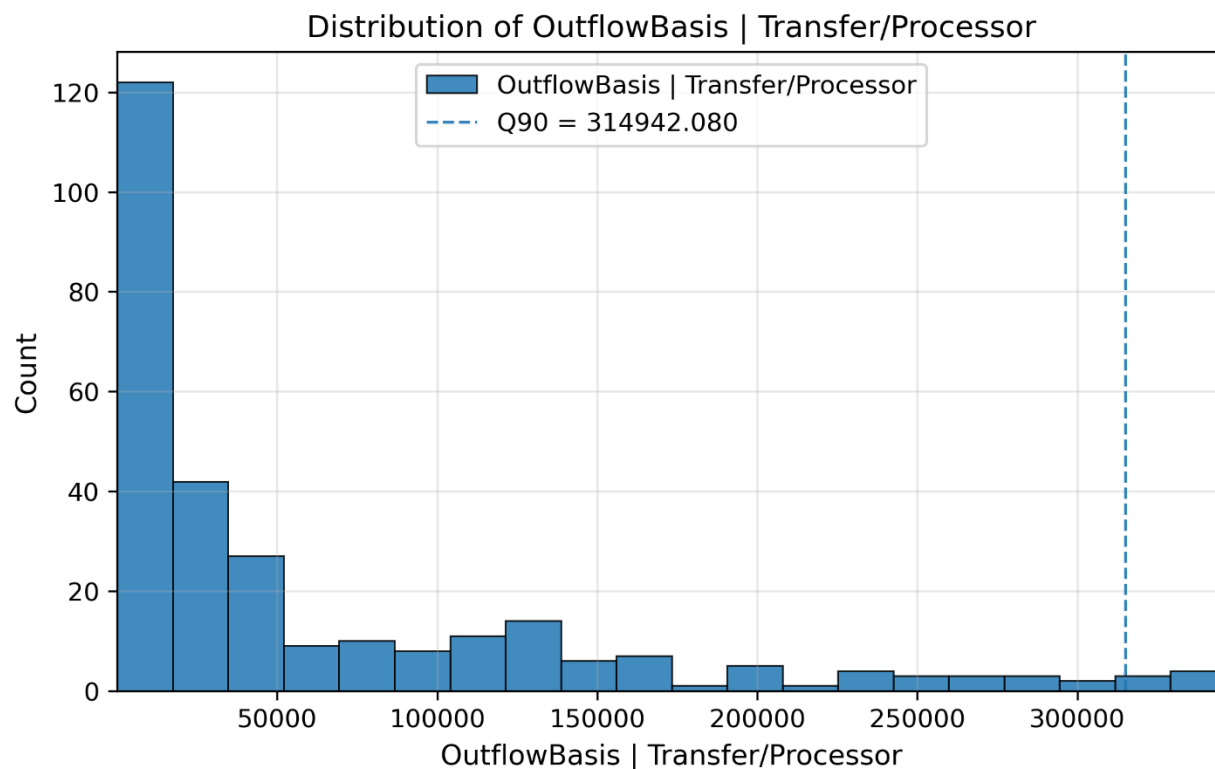


Figure 2.5 Histogram of Total Reported Outflows from Transfer/Processor Activities

This pattern is reinforced by the direct-to-disposal pathway. As discussed previously, approximately 16 million tons of waste are delivered directly to landfill facilities without passing through transfer processors, requiring supplementation from jurisdictional reporting to complete the disposal ledger (CalRecycle, 2023). Multimodality in Figure 2.6 demonstrates that there are several tiers of landfill size. More than half of the sites take in less than 150,000 tons per year. There is also an apparent mid-size group disposing of between 150 and 600 thousand tons and a relatively small number of landfills over 600,000 tons per year. Landfill disposal remains the dominant terminal pathway in the statewide system.

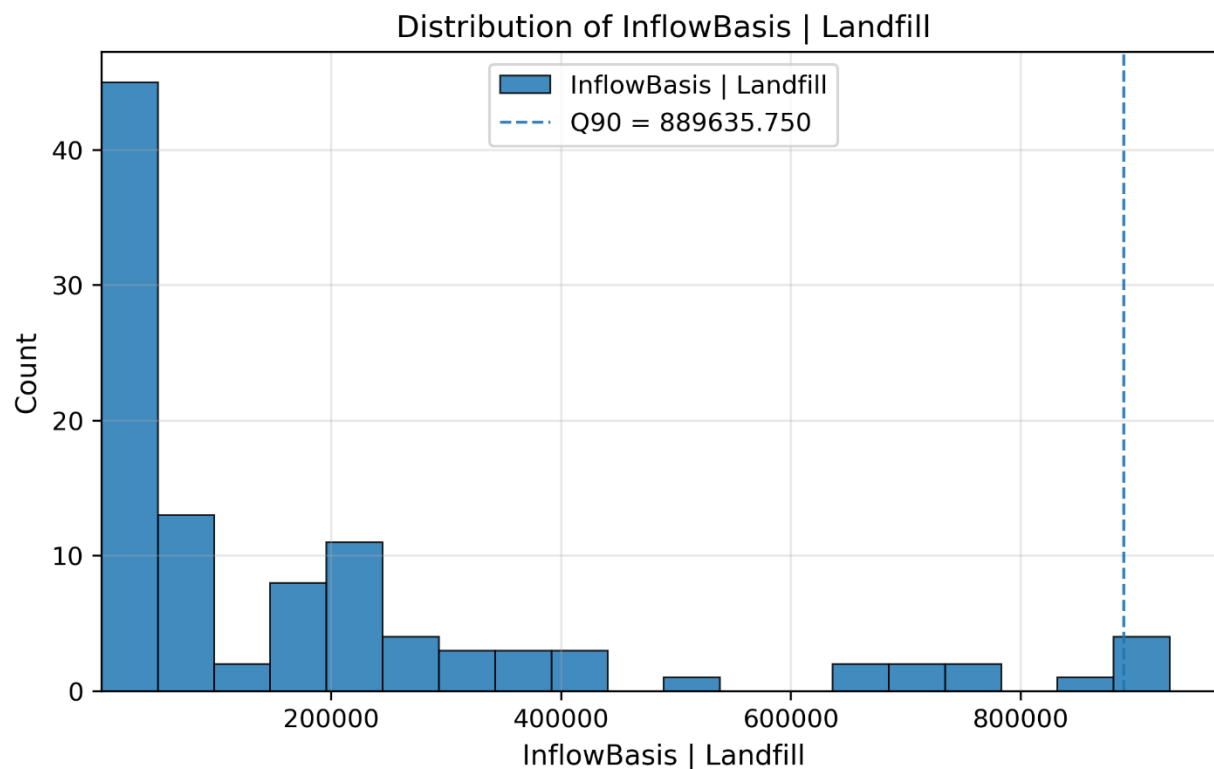


Figure 2.6 Histogram of Total Reported Inflows to Landfill Activities

The recycling/composting stream exhibits the greatest network complexity. Large volumes move not only from transfer/processors to recycler/composters (approximately 3.41 million tons) but also among similar activity types, including recycler-to-recycler flows (approximately 2.00 million tons) and transfer-to-transfer flows (approximately 2.11 million tons). These intra-class flows strongly suggest sequential processing stages, consolidation, contamination removal, commodity upgrading, and downstream specialization. Figure 2.7 shows the distribution of Outflows from Recycler/Composter Activities.

The end-use stream further reinforces the multi-stage structure of the system. While recycler/composter facilities account for the largest reported deliveries to end use, significant contributions also originate from transfer/processors and broker/transporter activities. This indicates that materials destined for final manufacturing, reuse, or market consumption may pass

through several intermediate reporting entities before reaching their terminal use, particularly in commodity chains involving fiber, CRV containers, metals, and processed organics.

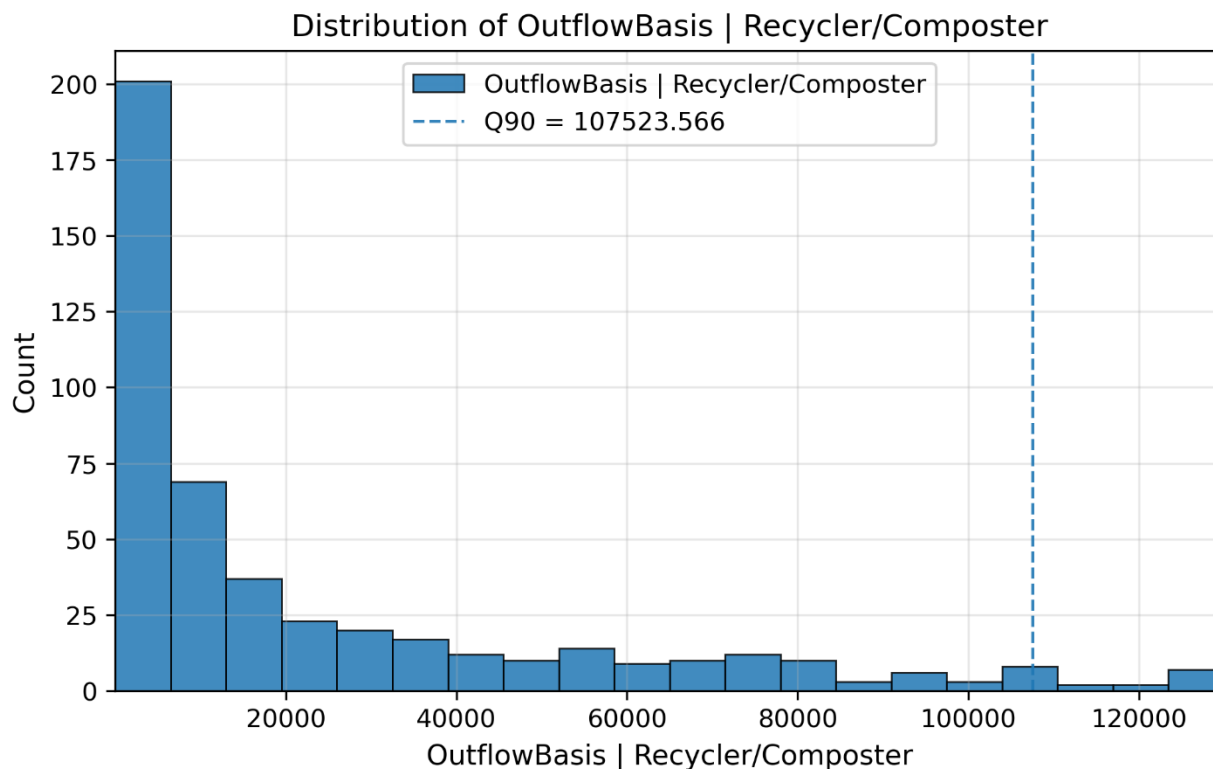


Figure 2.7 Histogram of Total Reported Outflows from Recycler/Composter Activities

Beneficial reuse flows are concentrated in movements from transfer/processors to landfills and from recycler/composters to landfills, historically corresponding to alternative daily cover (ADC), beneficial fill, and other reuse pathways that blur the distinction between disposal and diversion. Since implementation of SB 1383 requirements, the use of ADC as a diversion mechanism has been substantially restricted, reducing the long-term prominence of this pathway.

The brokering/transporting stream clearly captures intermediary logistics functions, with very large flows terminating at broker/transporter destinations, particularly from

recycler/composter and transfer/processor activities. This is consistent with market brokerage, commodity aggregation, and long-haul transport roles, especially for high-value recovered materials such as cardboard, metals, and certain plastic streams.

These aggregate patterns are further supported by the facility-level inflow and outflow histograms. The inflow distributions reveal strong right-skewness, with a relatively small number of facilities accounting for a disproportionate share of total throughput. This is especially pronounced for landfill and major transfer-processing activities, where large regional hubs dominate the upper quantiles. By contrast, recycler/composter and broker/transporter activities exhibit broader multimodal distributions, suggesting the coexistence of small, specialized facilities and large centralized processors.

The outflow histograms provide additional evidence of functional ambiguity. Landfills are characterized by extremely high inflows and minimal outflows, as expected for terminal disposal sites, whereas transfer/processors display the inverse pattern of comparatively modest reported inflows but very large outbound disposal flows. This asymmetry reflects the reporting structure of RDRS, in which collection-route deliveries are not comprehensively captured as inflows to transfer facilities. Recycler/composter activities often exhibit substantial outflows across multiple stream types, including both recyclable commodities and disposal residuals, further demonstrating that a single nominal activity class may encompass multiple latent operational roles.

Table 2.3 Material Outflows ('21 short tons) by Facility Activity

Material	MSW		CD		Organics	
Direction	To	From	To	From	To	From
Landfill	24,776,380	15,140	1,556,600	64,690	1,159,180	446,720

Transfer / Processor	384,550	24,741,850	399,230	2,656,710	972,700	4,349,070
Recycler / Composter	0	487,950	836,930	6,509,060	3,573,390	5,866,520
Broker / Transporter	36,910	242,580	33,190	85,040	772,610	1,039,700
Other	364,490	74,810	10,490	630	75,950	119,530
Unspecified	0	0	6,479,690	0	5,267,700	0
Material Total	25,562,330		9,316,130		11,821,530	
Material	Single Stream		Fiber		Metal	
Direction	To	From	To	From	To	From
Landfill	40,730	113,730	26,940	41,300	510	48,770
Transfer / Processor	1,218,420	1,189,910	361,510	1,268,010	86,120	370,590
Recycler / Composter	81,850	118,550	539,160	834,250	700,790	882,170
Broker / Transporter	250	23,810	1,455,420	1,569,320	131,350	24,980
Other	0	200	23,270	0	5,950	9,370
Unspecified	104,950	0	1,306,580	0	411,160	0
Material Total	1,446,200		3,712,880		1,335,880	
Material	Glass		Plastic		Other	
Direction	To	From	To	From	To	From
Landfill	70	11,250	1,140	5,180	752,240	35,490
Transfer / Processor	124,650	450,880	60,690	147,520	90,550	291,960
Recycler / Composter	479,340	747,540	170,940	379,430	83,790	752,020
Broker / Transporter	51,050	15,280	163,960	61,520	59,640	4,200
Other	0	0	3,910	0	1,280	79,940
Unspecified	569,830	0	193,000	0	176,120	0
Material Total	1,224,940		593,640		1,163,620	
Grand Total			56,177,150			

Taken together, the flow matrices and histogram distributions indicate that reported facility activity labels are administrative descriptors rather than unique process identifiers. The same nominal activity type may simultaneously participate in disposal, recovery, brokerage, and

end-use pathways. This empirical ambiguity directly motivates the methodological framework developed in Chapter 3, where facility classification is refined using observed flow interactions, material-specific routing patterns, and inferred operational functions rather than relying solely on reported activity categories.

2.6 Limitations of Material and Facility Typologies

The sections above establish a conceptual decomposition of California’s post-collection waste management system into material classes and nominal facility types and describe how these elements interact through empirically observed flows in the RDRS dataset. While this framework provides an essential foundation for understanding system structure, it is insufficient for uniquely characterizing facility operations or inferring process-level behavior. In practice, the combination of material categories and facility designations introduces several layers of ambiguity that limit interpretability and complicate system-level modeling.

First, material classifications in the reporting system are inherently coarse and do not uniquely determine physical composition or processing requirements. Broad categories such as municipal solid waste (MSW), organics, and construction and demolition (C&D) debris encompass highly heterogeneous mixtures of materials with distinct physical and chemical properties. For example, the “organics” category may include clean green waste, food waste, soiled paper, or mixed contaminated loads, each requiring different processing conditions and yielding different recovery outcomes. Similarly, MSW represents a residual category defined more by its disposal pathway than by its composition. As a result, the same reported material category may correspond to fundamentally different operational pathways depending on its underlying composition and contamination level.

Second, nominal facility type designations do not uniquely correspond to specific operational configurations. Regulatory labels such as “transfer/processing,” “recycler,” or “landfill” describe permitted activities rather than actual process-level behavior. Many facilities operate as hybrid systems, combining multiple functions within a single site. For instance, a transfer station may incorporate limited material recovery operations, while a landfill may co-locate sorting infrastructure or diversion programs. In such cases, reporting is often aggregated at the facility level, without disaggregation of intermediate processing steps or residual generation. Consequently, facilities with identical regulatory classifications may exhibit substantially different operational profiles, throughput characteristics, and recovery efficiencies.

Third, the structure of the RDRS dataset introduces additional ambiguity through incomplete material accounting and reporting conventions. The system records directional outflows from reporting entities but does not enforce a closed material balance. Inflows from non-reporting sources—such as direct collection routes or self-hauled waste—are not systematically captured, and downstream movements of recovered materials may fall outside reporting requirements once they are no longer classified as waste. This results in apparent “leakage” and “seepage” within the dataset, as visualized in Figure 2.8. Additionally, the presence

of unspecified destinations and aggregated reporting across multiple activities further obscures the interpretation of flows.

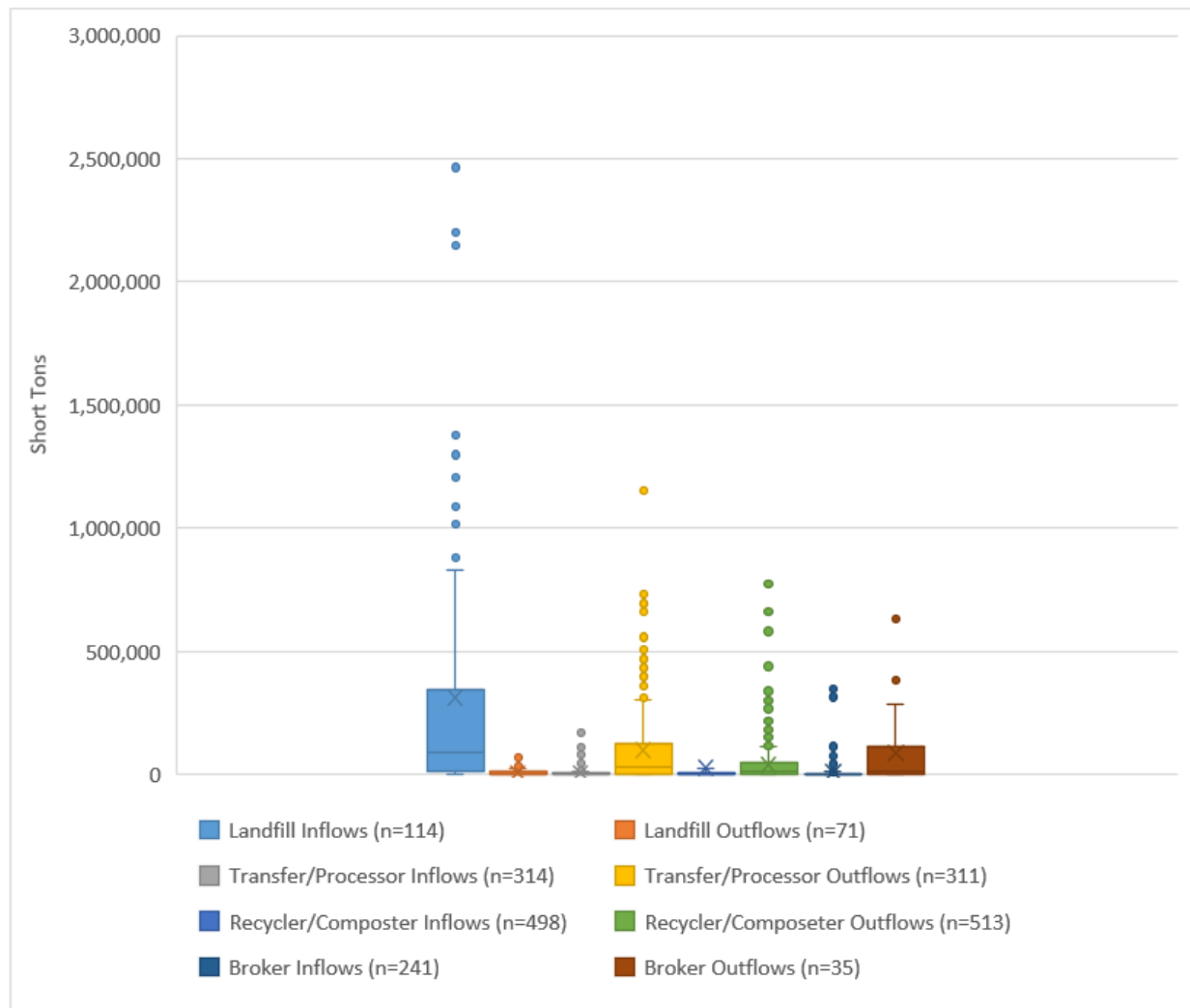


Figure 2.8 Box Plots of Facility-level Nonzero Inflows and Outflows

Figure 2.8 illustrates the systematic imbalances and the relative magnitude of the activities. When grouped in this way, the material ledger aggregates cannot adequately describe the categories much less explain the interactions between facilities. Ultimately, this

categorization of materials and facilities does not lead to a satisfactory descriptive model for any facilities besides landfills.

These three sources of ambiguity, interact to limit the explanatory power of the material–facility framework. While certain broad patterns can be reliably inferred (e.g., the role of transfer stations as logistical intermediaries or the movement of residual flows to landfills), the framework does not support detailed inference of process-level operations, recovery efficiencies, or emissions-generating activities. In particular, it is not possible to distinguish whether observed flows arise from sorting, preprocessing, residual disposal, or simple transloading without additional structural assumptions.

This limitation is not merely a data quality issue but reflects a deeper structural mismatch between regulatory classifications and the underlying physical processes of waste management. The categories used for reporting are designed for administrative and compliance purposes, rather than for representing the mechanistic transformation of materials through the system. As a result, the observed data must be reinterpreted through a more refined analytical framework that bridges this gap.

To address these limitations, Chapter 3 introduces a facility-level classification approach that decomposes the profile of observed material flows to infer operational roles. By moving beyond nominal facility labels, this framework enables a more precise representation of system behavior, supporting richer descriptive representation of facility interactions and estimation of emissions across the waste management network.

Chapter 3 Machine learning approach to identifying recycling and waste management facilities based on their inputs and outputs

This chapter develops a machine learning procedure for classification of facilities based on reported inflows and outflows. Section one introduces the research gap. Section 2 reviews relevant literature. Section 3 describes the dataset and the construction of facility-level material vectors. Section 4 outlines the methodological framework, including correspondence analysis, clustering, and consensus functions. Section 5 expounds the technical details of the implementation including calibration and application of methods, and heuristics. Section 6 presents the empirical results of the ensemble classification and interprets the identified facility archetypes. Section 7 concludes with a discussion of contributions, limitations, and directions for future research.

3.1 Introduction

Despite the importance of facility-level understanding, existing administrative classifications are ill-suited to capture the operational realities of waste infrastructure. The facility types recorded in CalRecycle’s Recycling and Disposal Reporting System (RDRS) and associated permitting databases reflect regulatory categories rather than functional archetypes. A single “Transfer/Processor” designation, for example, may encompass facilities that consolidate loads for long-haul transport, as well as facilities with sophisticated sorting equipment and recovery operations. Similarly, “Recycler/Composter” can refer to everything from specialized anaerobic digestion plants to open windrow composting sites as well as material recovery facilities (MRFs) and similar downstream processors. These ambiguities complicate the modeling of emissions, the evaluation of diversion strategies, and the broader goal of system integration. A data-driven reclassification of facilities, based on their observed material inflows and outflows, is therefore

needed to provide a better and operationally meaningful understanding of California's waste management infrastructure.

The underlying theoretical scaffold for the classification task is well-characterized as a two-stage process as shown in Figure 3.1 (Calvert and Young, 1974). A pattern vector is selected from available data to include the aspects and dimensions across which the modeler hopes to discriminate differences from one observation to the next. The pattern vector is first operated on by the feature extractor. This produces a feature vector which is reduced (into a subspace), transformed (into a feature space), or both compared to the original pattern vector. The resulting feature vector is then assessed by the classifier to assign a classification (grouping) under some optimizing framework. Most often, the general goal is to assign class membership to maximize the similarity between members within a certain class while minimizing the similarity between members of different classes.

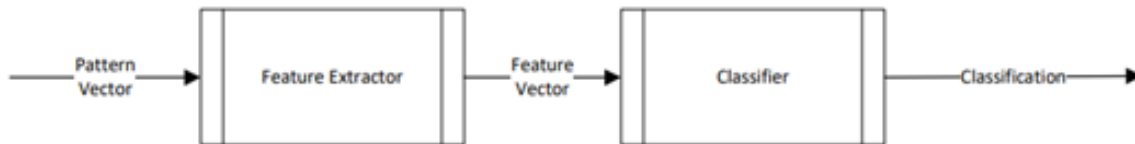


Figure 3.1 General pattern recognition paradigm

In this research, the implementation of pattern recognition is extended by implementing ensemble consensus classification. This involves calibrating multiple classifications of the same feature vector by different methods and parameterizations. The outputs of these runs are then used to form a new feature vector and a final class assignment which pools the results of the ensemble.

Strehl and Ghosh (2002) are widely regarded as the originators of the modern cluster ensemble framework, formalizing consensus clustering as a problem of knowledge reuse and defining the objective in terms of maximizing shared information across multiple partitions. Their formulation, grounded in normalized mutual information, established the principle that a collection of runs can coalesce into a knowledge base greater than the sum of its parts. Building on this foundation, Fern and Lin (2008) emphasize that a good ensemble must exhibit both quality and diversity. Quality refers to the extent to which individual ensemble members capture latent partition structure, subject to their method and parameters, whereas diversity describes the presence of complementary partitions that encode different but still plausible views of the data structure. In information-theoretic terms, a strong ensemble is one whose members collectively contain high mutual information within their partition structure while avoiding redundant solutions.

This formulation may be best represented as a graph. Figure 1 of (Fern and Brodley, 2004) provides a clear exposition of graph-based consensus representations through their bipartite formulation, in which the unit of analysis and cluster label are explicitly represented as distinct node sets (Instances and Clusters) connected by membership edges. This representation preserves membership information more succinctly than tabular data and provides a natural bridge to graph partitioning and community detection methods.

While this literature demonstrates the power of ensembles in abstract clustering problems, applications to real-world infrastructure data remain limited. Most experimental evaluations rely on benchmark datasets such as repositories, image retrieval tasks, or synthetic data designed to test algorithmic properties. These experiments provide methodological insights but rarely grapple with the messiness of administrative reporting, the sparsity of material flow

data, or the practical consequences of ambiguous facility types. Moreover, consensus clustering has not been widely used to interrogate the operational structures of waste management systems, where classification outcomes have immediate implications for emissions modeling, diversion strategies, and policy evaluation. From my research, there are no large-scale clustering treatments of this kind of data, and certainly not with the feature extraction cycle nor ensemble methods. This gap is especially salient in California, where facility-level reporting has only recently expanded under RDRS, offering a unique opportunity to apply ensemble consensus methods to a comprehensive dataset.

This chapter addresses these gaps by developing and implementing a data-driven classification of California's waste management facilities. Using the RDRS database of material flows, I construct facility-level input/output vectors harmonized under a custom schema of 16 material categories, including municipal solid waste, residuals, construction and demolition debris, organics, biosolids, and disaggregated recyclable commodities (fibers, metals, plastics, and glass). To enable interpretable clustering, I apply correspondence analysis (CoAn) to reduce dimensionality while preserving the chi-squared distances between material profiles. Multiple clustering algorithms and parameterizations are then applied to the transformed data, and ensemble consensus clustering is used to arrive at stable, descriptive and useful facility groupings. These consensus clusters are interpreted as operational archetypes that more accurately reflect the roles facilities play in California's integrated waste management system than existing administrative labels.

The contributions of this chapter are fourfold. First, it contextualizes the classification problem within the broader challenges of California's waste management and recycling system, emphasizing the policy and environmental stakes of facility-level understanding. Second, it

bridges the methodological literature on ensemble clustering with the applied problem of facility classification, showing how principles of quality, diversity, and consensus can be adapted to messy infrastructure data. Third, it provides a concrete implementation of ensemble consensus clustering on the RDRS dataset, demonstrating how correspondence analysis and clustering ensembles can work together to generate stable facility archetypes. Finally, it interprets the results in light of California’s integrated system, identifying both the strengths and limitations of the approach for informing emissions modeling and policy evaluation.

3.2 Literature Review: Classification and Integrated Waste Management Systems

This section reviews classification research as it relates to waste-management systems. It traces the intellectual and methodological lineage from early a priori typologies—used primarily for regulatory or descriptive purposes—to contemporary empirical, machine-learning, and ensemble-consensus frameworks capable of deriving operational classes directly from data. The review proceeds in four parts. Section 3.2.1 examines traditional and system-level classifications that organized waste infrastructure by process function, scale, and risk. Section 3.2.2 surveys the emergence of machine-learning and deep-learning approaches that automate material and facility identification across diverse data modalities. Section 3.2.3 synthesizes ensemble-clustering literature on stability and consensus, which provides the conceptual foundation for the multi-algorithm methodology adopted in this study. Finally, Section 3.2.4 broadens the perspective to include integrated and human–system interfaces—smart, IoT-enabled, and behavioral dimensions that connect algorithmic classification to real-world waste systems. Together, these literatures establish the theoretical and empirical basis for the ensemble-consensus framework implemented in this work.

3.2.1 Empirical and System-Level Classification in Waste Systems

The following section reviews how classification has evolved within the waste-management literature itself. Early studies emphasized administrative or engineering typologies designed for policy compliance and reporting rather than for analytical modeling. Subsequent research introduced empirical and system-based approaches, using measured flows, risks, and treatment processes to derive operational facility classes. This evolution from descriptive to quantitative frameworks forms the foundation for understanding how data-driven methods can later extend these classifications.

Early approaches to waste system classification were rooted in a priori typologies that reflected administrative or engineering practice rather than empirical data. Demirbaş (2011) provided one of the most comprehensive summaries of these frameworks, defining waste management as a sequence of generation, collection, treatment, and final disposition. His classification divided sources into domestic, commercial, industrial, biomedical, construction, and hazardous wastes, emphasizing that each category implied characteristic handling and processing technologies. The resulting “waste hierarchy” (reduce, reuse, recycle, recover, dispose) remains foundational in planning documents and informs the categorical facility codes used by agencies such as CalRecycle and the U.S. EPA. Although conceptually tidy, this rule-based typology presumes clear boundaries between sources and processes and does not easily accommodate facilities that combine multiple roles (e.g., hybrid transfer–MRF operations).

A second line of research translated these conceptual typologies into quantitative system models that treat classification as a functional descriptor of flow and performance. Wang and You (2021) developed a system-dynamics model for municipal solid waste in Tianjin, China, in which classification of MSW by treatment type (source-separated, mechanically sorted, or

mixed) directly affected model outcomes such as greenhouse-gas reductions and economic returns. This perspective aligns closely with the present study's statewide scope, in which facility classifications influence modeled transport, energy use, and emissions outcomes.

Further refinement occurred as researchers sought to differentiate facilities by risk and performance characteristics rather than by waste type alone. Nišić and Knežević (2018) introduced a risk-based classification for industrial waste-storage facilities, integrating hazard potential, containment quality, and environmental sensitivity into a quantitative scoring framework. Such frameworks anticipate modern regulatory trends emphasizing adaptive management and data transparency.

The evolution from typological to empirical and then to technologically mediated classification is captured in Yang et al. (2021), whose bibliometric review (2000–2019) shows a clear shift from mechanical sorting methods to optical, spectroscopic, and image-recognition techniques. They mark two research phases, property-based and identification-based, which parallels the progression toward AI-driven recognition.

Finally, Davis et al (2021) demonstrated the first fully operational deep convolutional neural network (VGG-16) for classifying construction and demolition waste, achieving ~94 % accuracy across seven material classes. The VGG-16 architecture essentially introduces more layers (depth) but applies smaller convolutions (analogous to reducing the calculation window for a time-series moving average). Their study exemplifies the extension of empirical classification to data-driven automation.

Collectively, these studies trace the methodological evolution from conceptual typologies to quantitative and data-empirical classification, establishing the intellectual bridge to contemporary ML frameworks.

3.2.2 Data-Driven, Machine-Learning, and Deep-Learning Classifications

While these empirical frameworks grounded facility classification in measurable flows and performance indicators, they remained largely deterministic and system-specific. The rise of machine learning and artificial intelligence has since transformed this landscape, enabling classifications derived directly from data rather than predefined categories (Ahmad et al., 2020; Chu et al., 2018; Li and Chen, 2023; Lin et al., 2024; Nahiduzzaman et al., 2025).

The past decade has witnessed a decisive methodological turn from descriptive or rule-based classification toward machine-learning (ML) and deep-learning (DL) frameworks that infer categories directly from data. Chu et al. (2018) pioneered a multilayer hybrid system combining convolutional neural networks (CNN), multilayer perceptrons, and auxiliary sensors to classify municipal waste. Ahmad et al. (2020) advanced this through an intelligent double-fusion strategy that optimally combined multiple deep architectures including AlexNet, VGG, GoogleNet, and ResNet, using feature- and score-level fusion with particle-swarm and genetic optimization, improving accuracy by $\approx 3.6\%$. Li and Chen (2023) embedded CNN feature extraction within a Graph-LSTM to model contextual dependencies among objects in municipal solid waste, achieving 97.5 % accuracy (against repositories) and demonstrating the power of relational representation. Lin et al. (2024) proposed a multi-scale feature-fusion strategy (MFFS) combining VGG16 and Vision Transformer modules for edge deployment, achieving 95 % accuracy with 57% fewer parameters. Finally, Nahiduzzaman et al. (2025) extended deep-learning classification to IoT-integrated urban systems, embedding neural classifiers within

smart-collection networks for real-time route optimization. These are very different problems than the facility classification problem in my research.

Across these studies, two methodological principles emerge:

1. **Complementarity and Consensus:** Accuracy improves when multiple models or feature views are fused rather than used independently.
2. **Scalability and Context Awareness:** Modern models encode relationships among waste items, processes, or system nodes instead of treating samples as independent.

These principles directly motivate the ensemble-consensus approach implemented in this dissertation, which transfers the logic of deep-fusion models to large-scale, heterogeneous facility data. Furthermore, these are supervised learning whereas, my method's individual steps are unsupervised. And in connecting them I seek, across all levels of the feature extraction(s) and clustering method(s), automating rules to bridge parameter-setting. This philosophy brought me to ensemble clustering

3.2.3 Consensus and Stability in Ensemble Clustering

Despite the success of deep learning, unsupervised classification remains prone to instability: different parameterizations or random seeds may produce divergent partitions. This issue is especially consequential for waste-facility data, where variability in reporting, scale, and composition introduces sensitivity to modeling choices.

Strehl and Ghosh (2002) first formalized ensemble clustering as the combination of multiple partitions into a unified labeling that maximizes overall agreement. They introduced the co-association matrix, representing how often pairs of objects co-occur across clustering runs, and demonstrated that consensus partitions can be extracted from this meta-similarity graph.

(Fern and Lin, 2008) expanded this framework by analyzing the trade-off between cluster quality and diversity, showing that structured diversity, generated through feature randomization or algorithmic variation, enhances consensus robustness. Feffer et al. (2022) later extended ensemble clustering into graph-theoretic and community-detection domains, aggregating adjacency matrices from multiple runs to identify stable node communities.

In this dissertation, these principles guide the implementation of ensemble consensus to derive stable facility archetypes. Multiple clustering methods—k-means, hierarchical, and mean-shift—generate candidate partitions of the correspondence-analysis feature space. Their co-association frequencies are then used to build a consensus graph, from which the final stable classification is obtained. This structure embodies Strehl and Ghosh’s meta-clustering logic, Fern & Lin’s balance of quality and diversity, and Feffer et al.’s graph-based integration, ensuring the resulting typology is robust, reproducible, and interpretable.

3.2.4 Emerging Integrations and Human–System Interfaces in Waste Classification

Ensemble consensus techniques establish a methodological bridge between algorithmic diversity and empirical robustness. Yet classification research is expanding beyond algorithms toward integration with sensing networks, behavioral factors, and policy contexts.

Smart and IoT-Enabled Systems: Recent studies situate classification within IoT and smart-city infrastructures, employing distributed sensors and embedded computing to enable real-time adaptive waste management (Ahmed et al., 2023; Cheema et al., 2022; Rahman et al., 2020; Wang et al., 2021). These systems couple visual or ultrasonic sensors with lightweight CNNs deployed on edge devices, feeding live data to optimize collection routes and facility scheduling.

Behavioral and Policy Dimensions: A complementary vein of research examines human participation and policy design in supporting classification accuracy. Studies (Zhang et al., 2022; Zhang and Huang, 2021) show that resident engagement, education programs, and incentive mechanisms (e.g., deposit-refund schemes) significantly improve the consistency of waste separation, indirectly strengthening data quality for ML models.

Sector-Specific and Specialized Applications: Emerging research extends ML classification to specialized domains such as medical waste (Zhou et al., 2022) and construction materials (Li et al., 2023), where data imbalance and safety constraints drive methodological innovation through transfer learning and class-balanced loss functions.

Review and Meta-Analysis Literature: Recent meta-analyses (Nasir and Al-Talib, 2023; Toğaçar et al., 2020) synthesize trends across hundreds of studies, identifying convergence around lightweight architectures, robotic integration, and multi-modal fusion. They also note persistent gaps in dataset standardization and reproducibility, motivating the need for broader, system-scale approaches.

In summary, these works reveal a clear progression—from normative typologies, through empirical and machine-learning approaches, to ensemble-consensus and integrated intelligent systems. Despite this progress, a critical gap remains at the system scale: few studies combine diverse data sources, algorithmic ensembles, and facility-network representations to derive stable operational archetypes. The methodology that follows addresses this gap by applying ensemble-consensus principles to statewide waste-flow data, enabling reproducible facility classification across heterogeneous and incomplete information environments.

3.3 Data Description and Pattern Specification

The dataset is drawn from CalRecycle’s RDRS reporting tool, which records material movements between waste-handling facilities in California (CalRecycle, 2026m). Quarterly reports are made to CalRecycle containing information on material type and tonnage as well as the destination facility. The RDRS marks a recent expansion of this reporting to include a much wider range of materials and facility types. Before 2019, only transfer stations and landfills were required to report movements of solid waste or alternative cover in the DRS system (CalRecycle, 2026n). CalRecycle kindly gave us access to the full reporting of outflows for all facilities and materials in RDRS.

Unfortunately, the recorded facility types are not well-defined and coincide with regulatory, legal, or administrative categories rather than operational archetypes. For example, the ambiguity and breadth of the categories “Transfer/Processor” and “Recycler/Composter” necessitate reclassification. In the context of air emissions modeling, there should be very different mechanisms of pollutant release based on the degree of heavy machinery usage, electrical power consumed, and associated process emissions. Even among “Landfills” there are significant operational variation due to the colocation and combined reporting of multiple activities (MRF, composting, etc.) occurring quite literally on top of some landfills. Thus, the goal of the classification is to better distinguish operational profiles, inferred through close examination of reported material movements.

First, the reported materials were harmonized under a custom schema to reduce noise from idiosyncratic entries and emphasize materials relevant to my modeling goals. After several iterations, the final version used in the classification includes 16 material categories. This includes the three most common materials, Municipal Solid Waste (MSW), Construction and

Demolition Debris (CD), and Organics. A “residuals” category (meaning contamination removed during processing of other streams) is included alongside MSW. Organics are divided into Biosolids (a narrowly defined process output of water reclamation plants) and all other organics (including food waste, garden waste and similar waste streams). Recyclables are separated into nine categories representing four broad groups: Fibers consists of Paper and Cardboard, Metals consists of Ferrous, NonFerrous, and Mixed Metal, Plastics consists of PET, HDPE, and Mixed Plastic, and Glass is reduced to a monolith. Single Stream is a mixed recycling stream, often

Table 3.1 Material categories in RDRS and reported tons of 2021 outflows

Description	Category	Reported Outflows (short tons)
Waste	MSW	25,562,337
	Residuals	695,984
CD	CD	9,316,128
Organics	Organics	8,368,205
	Biosolids	3,453,330
Mixed Recyclables	SingleStream	1,446,199
Fiber	Cardboard	2,354,667
	Paper	1,358,214
Metal	Ferrous	699,584
	NonFerrous	293,944
	Mixed Metal	342,357
Plastic	PET	305,894
	HDPE	115,728
	Mixed Plastic	172,031
Glass	Glass	1,224,944
Other	Other	467,629
TOTAL	TOTAL	30,614,838

used to indicate residential or commercial mixed recycling “as discarded” – that is, nominally recyclable but likely contaminated and requiring further processing. Bulky items such as appliances, furniture, carpet, mattresses and tires are assigned to “Other” – in California, these all have targeted diversion programs (CalRecycle, 2025). Further subdivision of these categories is possible but the quantities become quite small, introducing even more zeroes to what is already a sparsely-populated pattern vector. Table 3.1 indicates reported totals for each of the materials in the schema.

Second, the database was recast as a ledger of annual inflows and outflows for each facility, aggregated by the material schema. Due to the nature of the database and reporting modalities of certain waste operations, there is no expectation of a net-zero material balance. To be more specific, the database records outflows reported by the origin facility – this means that waste flows from locations which are not registered reporting entities (e.g., homes or businesses) are not reported explicitly. For instance, most waste appears at transfer stations as an outflow from transfer stations to landfill. Similarly, waste flows into landfills but not out – we can assume that it is disposed at that location.

This results in the final pattern vector, a 32-dimensional pattern vector (inbound and outbound quantities for each of 16 material categories). The combination of inbound and outbound components was motivated by the need to infer activities as changes in material designations (e.g. sorting single stream into separated recyclable commodities), thereby encapsulating operational behavior in a single vector.

To enhance the interpretability of my analysis, a series of data filtering and cleaning steps were applied. Facilities were excluded if no inflows or outflows are reported in 2021, as such entities either failed to report relevant tonnage or had negligible activity for the year in question.

California regulations require reporting if there is more than 100 tons of material handled per quarter. For each retained facility, the material vector was normalized to a row profile, that is, each vector was scaled such that its elements represented the proportion of each material type relative to the facility's total throughput. This normalization was critical for enabling a meaningful application of the correspondence analysis, which is sensitive to compositional variation but agnostic to scale. Row profiles ensure that differences between facilities reflect changes in material composition rather than differences in total volume processed. After the transformation, I append a standard log-normal measure of facility size.

Ultimately, for year 2021, there are 1,683 observations each represented by a 32-dimensional pattern plus identifying keys.

3.4 Classification Methodology

In the following section, I explain the mathematical underpinnings of the methods of transformation, feature extraction, and clustering utilized in my implementation of ensemble clustering.

3.4.1 Feature Extraction by Correspondence Analysis

In the conceptual framework described above, the feature extractor operates on the pattern vector to produce the feature vector. The goal of this step is to enhance the important aspects of the pattern, reduce noise, and often includes transformation of the original pattern into similarity or distance measures for use in the subsequent classifier. These techniques can be quite simple, such as applying a scale factor to normalize a variable. An intermediate application would include standardization of a variable because it involves examining all observations to calculate the mean and variance. I will utilize Correspondence Analysis as an analytical method of

transformation and dimensionality reduction designed to decrease the computational burden of clustering while sacrificing as little explanatory power as possible.

Correspondence analysis (CoAn) provides a means of generating an optimally scaled and rotated representation of the original pattern vector (Greenacre, 2007). As with principal component analysis (PCA), a decomposition of the rank and eigenstructure of the matrix is performed (Anderberg, 1973), allowing for statistical analytical dimensionality reduction. The distinction lies in what is being decomposed. PCA operates on continuous, quantitative data using Euclidean distance, seeking to maximize explained variance across orthogonal axes. CoAn, on the other hand, is designed for categorical count data arranged in a contingency table. This makes it suitable for all kinds of data as a qualitative multivariate statistical method (Golob, 1985). It decomposes the chi-squared distances between rows and columns, preserving relational structure while enabling projection into a Euclidean subspace. Thus, CoAn summarizes the association patterns in categorical data, where distances reflect differences in distributional profiles, and the goal is to capture departures from independence.

In correspondence analysis, each row of a contingency table is expressed as a row profile, a vector of relative frequencies across columns. The chi-squared distance between two profiles is the weighted sum of squared differences in these relative frequencies, with weights equal to the inverse of column proportions. This formulation gives higher leverage to rare categories, emphasizing differences where they matter most in terms of statistical association. The result is that CoAn measures similarity of profiles rather than absolute frequency. Two rows may be considered close even if their totals differ greatly, as long as their proportional distributions across columns are similar.

Geometrically, CoAn maps these chi-squared distances into a Euclidean space. The positions of rows and columns on the resulting biplot approximate their relationships: categories that are close together share similar patterns of distribution, while those far apart differ strongly. The total inertia in CoAn, analogous to variance in PCA, is proportional to the Pearson chi-squared statistic of independence. This direct link means that the post-transformation geometry represents the extent to which the joint distribution of rows and columns departs from statistical independence.

The following calculations characterize the transformation:

- Let $\mathbf{N}$ denote the initial $I \times J$ matrix of data (nonnegative elements, column and row sums are positive).
- Calculate the correspondence matrix profiles ($\mathbf{P}$) by dividing each cell by the grand total.
- Compute the inverse square roots of the diagonalized row ($\mathbf{r}$) and column ($\mathbf{c}$) sums of $\mathbf{P}$, denoted $\mathbf{D}_r^{-\frac{1}{2}}$ and $\mathbf{D}_c^{-\frac{1}{2}}$
- Then, calculate the matrix of standardized residuals $\mathbf{S}$:

$$\mathbf{S} = \mathbf{D}_r^{-\frac{1}{2}}(\mathbf{P} - \mathbf{r}\mathbf{c}^T)\mathbf{D}_c^{-\frac{1}{2}} \quad (3.1)$$

- And finally, calculate the Singular Value Decomposition (SVD) of $\mathbf{S}$. This finds $\mathbf{U}$, $\mathbf{V}$, and $\mathbf{D}_\alpha$ satisfying:

$$\mathbf{S} = \mathbf{U}\mathbf{D}_\alpha\mathbf{V}^T \quad \text{s. t.} \quad \mathbf{U}^T\mathbf{U} = \mathbf{V}^T\mathbf{V} = \mathbf{I} \quad (3.2)$$

The transform is applied by the calculations indicated in Table 3.2. Typically, principal rows and standard columns are charted in a so-called biplot together or vice versa.

Table 3.2 Equations to apply the estimated CoAn transform to Row- or Column-profiles

Coordinate Systems	Rows	Columns
Standard	$\Phi = \mathbf{D}_r^{-\frac{1}{2}} \mathbf{U} \quad (3.3)$	$\Gamma = \mathbf{D}_c^{-\frac{1}{2}} \mathbf{V} \quad (3.4)$
Principal	$\mathbf{F} = \mathbf{D}_r^{-\frac{1}{2}} \mathbf{U} \mathbf{D}_\alpha \quad (3.5)$	$\mathbf{G} = \mathbf{D}_c^{-\frac{1}{2}} \mathbf{V} \mathbf{D}_\alpha \quad (3.6)$

From a conceptual point of view, CoAn is ideal for this application. The hypothesis is that the operations of a given facility are revealed by the particular materials entering and leaving. Moreover, with a goal of classifying types of facilities, the profile (composition) of the material-handling footprint is more important than the size of the operation. Another benefit of this method is the resulting projections map each facility into a Euclidean space where facilities with similar material distributions cluster along similar trajectories, independent of scale. The transformed profiles are amenable to various similarity measurement calculations.

Based on the result of the CoAn transformation, approximation of the system through dimensionality reduction is achieved by retaining a subset of the highest singular values. These dimensions can be said to account for a quantifiable portion of the total observed variation or inertia. Ideally, the number of dimensions to achieve a significant fraction of explained variation is small - these terms, “significant” and “small” are, of course, subjective.

3.4.2 Cosine Similarity Measure

Viewing the transformed profiles as vectors emanating from the origin of the Euclidean feature space leads naturally to the interpretation of their angles as a similarity measure. Cosine

similarity, defined as the normalized dot product of vectors, measures angular similarity and is appropriate in high-dimensional spaces where magnitude varies but directional patterns remain informative. Use of the cosine of the angle subtended between each vector (Koppelman and Pas, 1985) as a measure of similarity produces a consistent measure of identical range for each of the component dimensions. The measure is scalable in terms of both the dimensionality of the vectors being compared as well as the number of comparisons being made per observation.

Each facility's feature vector, as projected in CoAn-transformed Euclidean space, has its cosine similarity calculated relative to a set of reference vectors. This makes the choice of reference vectors important. Letting the cosine be evaluated pairwise for all observations would lend itself to hierarchical clustering and making an arbitrary selection of one reference vector could seriously bias results or hinder the effectiveness of subsequent clustering. I adopt a strategic approach that selects empirical profiles as reference profiles by maximizing angular dispersion. Specifically, adjacent pairs of retained principal dimensions (1–2, 2–3, etc.) are used to identify facility vectors exhibiting maximal orthogonality, and a final vector is chosen for global orthogonality across all retained dimensions. As a rule, the number of reference vectors is selected as the number of principal dimensions retained. This design ensures wide angular coverage, avoids concentration near any single archetype and improves the discriminatory power of cosine similarity.

3.4.3 Log transform

While the CoAn input matrix is scale-invariant, it is also desirable to retain a scalar summary of facility magnitude for optional downstream use. To this end, the facility basis was defined as the sum of all inbound and outbound tonnage for each facility prior to normalization. Because the distribution of basis values is highly skewed, spanning multiple orders of magnitude, I applied a

log-transformation using the function $\log(x + 1)$ which compresses the long tail of large facilities while preserving ordering and relative differences. The log-transformed basis was then standardized to a z-score, resulting in a normalized variable that captures size variation in standard deviation units. This log-z facility size basis is appended alongside cosine-based similarity features for inclusion in classification procedures.

3.4.4 Classification Methodologies

Multiple Classification methodologies are utilized in this research. These are k-means, hierarchical, mean shift, and spectral clustering. Furthermore, the concept of ensemble consensus clustering is explained.

3.4.4.1 K-Means Clustering:

Given a dataset $\mathbf{X} = \{\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_n\} \in \mathbf{R}^d$, the K-Means algorithm seeks to partition the data into k clusters $C_1, C_2, \dots, C_k$ such that the within-cluster variance is minimized (Everitt, 1993).

Formally, it solves the following optimization problem:

$$\min_{C_1, \dots, C_k} \sum_{j=1}^k \sum_{\mathbf{x}_i \in C_j} \|\mathbf{x}_i - \boldsymbol{\mu}_j\|^2 \quad (3.7)$$

where $\boldsymbol{\mu}_j$ is the centroid of cluster C_j , defined as the mean of all points in the cluster:

The algorithm proceeds iteratively via two main steps:

1. **Assignment step:** Each point is assigned to the cluster with the nearest centroid.
2. **Update step:** Centroids are recomputed as the mean of the points assigned to each cluster.

These steps are repeated until the assignments converge or the change in objective function is below a predefined threshold.

To determine the most appropriate number of clusters k , two key metrics are used: inertia and silhouette score. These metrics provide complementary insights into the quality and stability of the clustering solution.

3.4.4.2 Agglomerative Hierarchical Clustering

For any two facility vectors $\mathbf{z}_i$ and $\mathbf{z}_j$ of dimension M , $d(\mathbf{z}_i, \mathbf{z}_j)$ denotes the Euclidean distance between $\mathbf{z}_i$ and $\mathbf{z}_j$. This metric is used in both the clustering algorithm and the silhouette score computation.

At each step of the agglomerative process, two clusters C_a and C_b are merged based on the average linkage criterion (Calvert and Young, 1974; Everitt, 1993):

$$D(C_a, C_b) = \frac{1}{|C_a| \cdot |C_b|} \sum_{\mathbf{z}_i \in C_a} \sum_{\mathbf{z}_j \in C_b} d(\mathbf{z}_i, \mathbf{z}_j) \quad (3.8)$$

This criterion balances intra-cluster cohesion and inter-cluster separation and avoids the bias of extreme distances found in complete or single linkage. The silhouette score is computed to evaluate clustering quality.

The output of agglomerative clustering is a dendrogram, a tree structure encoding merge operations. Each node represents a cluster, and each merge is associated with the inter-cluster distance $D(C_a, C_b)$. The height of each merge in the dendrogram reflects the distance at which clusters were joined, offering an interpretable view of nested similarity.

3.4.4.3 Mean-Shift Clustering

The mean-shift clustering algorithm operates by iteratively shifting each data point toward the center of mass of its local neighborhood, effectively performing gradient ascent in a smoothed kernel density estimate of the data. Let $\mathbf{X} = \{\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_n\} \in \mathbf{R}^d$ denote the set of facility feature vectors, composed of cosine similarity values to reference profiles and a standardized log-transformed size variable. The update rule for a data point $\mathbf{x}_i$ at iteration t is given by:

$$\mathbf{x}_i^{(t+1)} = \frac{\sum_{j=1}^n K\left(\frac{\|\mathbf{x}_j - \mathbf{x}_i^{(t)}\|^2}{h^2}\right) \mathbf{x}_j}{\sum_{j=1}^n K\left(\frac{\|\mathbf{x}_j - \mathbf{x}_i^{(t)}\|^2}{h^2}\right)} \quad (3.9)$$

where h is the bandwidth parameter that controls the radius of influence for the kernel K , often Gaussian. After convergence, all points that move toward the same mode are assigned to the same cluster.

The bandwidth parameter plays a critical role in the behavior of Mean Shift. Intuitively, it determines the granularity at which modes are detected: a small bandwidth detects fine-grained, potentially noisy distinctions, while a large bandwidth smooths over local variability to emphasize broader patterns. In scikit-learn, the `estimate_bandwidth()` function provides a data-driven heuristic for bandwidth selection by computing the pairwise distances between a sample of points and selecting a user-specified quantile as the effective radius. Although this approach offers a starting point, the algorithm's sensitivity to small changes in bandwidth—even when the number of clusters remains the same—necessitates a careful evaluation of each result.

To explore the effect of the bandwidth parameter, a critical factor controlling cluster granularity, clustering was performed across a range of bandwidth values derived from quantile-

based distance estimation. The quantile is referencing the distribution of inter-observation distances, similar to the linkage distance in hierarchical clustering. A small quantile corresponds to a lower bandwidth and more clusters result and vice versa. The explored quantiles differ by experiment, determined empirically by a screening run – bounds are set from the lowest quantile yielding a number of clusters less than 40 up to the lowest quantile yielding 2 clusters. This range is then explored at intervals of 0.05.

The resulting cluster solutions were evaluated using silhouette scores and the stability of certain cluster granularities. In terms of stability, a property of mean shift clustering is that multiple values of the bandwidth parameter (a continuous variable) map to the discrete, number of clusters outcome. In general, there is a range of bandwidth that maps to a single value for a number of clusters. However, the same number of clusters does not necessarily mean the same membership results. Furthermore, due to some randomness in the selection of initial clusters, the ranges are not strictly closed. That is, a small increase [decrease] in quantile/bandwidth can result in an increase [decrease] in the number of clusters, albeit rarely.

This approach allows Mean Shift clustering to function as a flexible, exploratory tool that adapts to the density structure of the data, offering insight into the latent groupings of facilities based on both material profile similarity and throughput scale. Its integration with cosine-transformed features further supports its use in identifying facilities that are behaviorally similar, even when their magnitude of operations differs.

3.4.4.4 Spectral Clustering

Spectral clustering is a graph-theoretic clustering technique that leverages the eigenstructure of a similarity matrix to identify natural partitions in data. Unlike traditional centroid-based approaches such as K-means, spectral clustering does not assume spherical clusters or linear separability. Instead, it interprets data as a graph where each observation is a node, and edges between nodes are weighted by a similarity measure (e.g., Gaussian kernel, cosine similarity, or co-association frequency). Clustering is achieved by embedding the graph in a lower-dimensional spectral space using the leading eigenvectors of the graph Laplacian (see below), followed by a standard clustering algorithm (usually K-means) in that space.

The mathematical foundation of this approach involves constructing a graph Laplacian $L = D - W$, where W is the similarity (affinity) matrix and D is the diagonal degree matrix with entries $D_{ii} = \sum_j W_{ij}$. In its normalized form, the Laplacian becomes $L_{sym} = I - D^{-1/2} W D^{-1/2}$. The top k eigenvectors of L_{sym} capture the intrinsic structure of the data – this defaults to the number of clusters if no alternative is provided. Each row of the resulting spectral embedding matrix corresponds to an observation and encodes its position in the feature subspace. These embedded vectors are then grouped using a simple clustering algorithm such as K-means.

Spectral clustering is particularly advantageous when dealing with complex cluster shapes, disconnected components, or manifold-structured data. It performs well when cluster boundaries are not convex and is relatively robust to noise when the similarity matrix is constructed appropriately. Its flexibility in accepting arbitrary similarity matrices also makes it compatible with ensemble-based similarity approaches such as co-association matrices.

In addition to standard internal validation metrics such as the silhouette score, spectral clustering offers a unique model selection heuristic based on the eigengap of the graph Laplacian.

3.4.4.5 Internal Validation Measures for Selected Cluster Methods

Inertia (Within-Cluster Sum of Squares)

Inertia is the objective function minimized by K-Means. It measures the compactness of clusters by summing the squared distances between each point and its corresponding cluster centroid:

$$\text{Inertia} = \sum_{j=1}^k \sum_{\mathbf{x}_i \in C_j} \|\mathbf{x}_i - \boldsymbol{\mu}_j\|^2 \quad (3.10)$$

Lower values of inertia indicate tighter clusters. However, inertia always decreases with increasing k , even when additional clusters are not meaningful. Therefore, inertia alone is not sufficient for selecting k , but it is useful for identifying an “elbow” point where additional clusters offer diminishing returns.

Silhouette Score

The silhouette score offers a normalized assessment of how well each observation fits within its assigned cluster, relative to other clusters (Shahapure and Nicholas, 2020). For each data point $\mathbf{x}_i$, the silhouette value s_i is defined as:

$$s_i = \frac{b_i - a_i}{\max(a_i, b_i)} \quad (3.11)$$

where:

- a_i is the average distance from $\mathbf{x}_i$ to all other points in the same cluster (intra-cluster cohesion).

- b_i is the minimum average distance from $\mathbf{x}_i$ to all points in the nearest different cluster (inter-cluster separation).

The global silhouette score S is the average of s_i over all points. Values close to 1 indicate well-separated, compact clusters; values near 0 indicate overlapping clusters; negative values suggest potential misclassification. The overall silhouette score is the average of s_i over all data points. It is an effective criterion for comparing different values of k : an upward jump in silhouette score usually suggests a more natural clustering structure, but it also naturally increases slightly with the addition of clusters.

Eigengap

The eigengap is defined as the difference between consecutive eigenvalues of the Laplacian matrix, and it provides insight into the natural number of clusters inherent in the data's similarity structure. Specifically, if the normalized Laplacian is used, its eigenvalues capture the connectivity and structure of the similarity graph derived from the data (van Dam et al., 2021). A large gap between the k^{th} and $(k + 1)^{th}$ smallest eigenvalues indicates that a partition into k well-separated clusters may be optimal. This is motivated by the theoretical interpretation of spectral clustering as a relaxed solution to the normalized cut problem in graph partitioning. In practice, the eigengap heuristic is visualized by plotting the sorted eigenvalues and identifying the largest jump, which suggests the most stable number of clusters

3.4.5 Ensemble Consensus Clustering

The co-association matrix is a key construct in ensemble clustering, which aims to combine multiple clustering results into a single, more robust consensus solution. The co-association

matrix is an $N \times N$ symmetric matrix (where N is the number of observations, 1329 in my final setup) that quantifies how often each pair of observations is grouped together across different clustering runs or algorithms. For any two observations i and j , the entry C_{ij} represents the proportion of clustering results in which they appear in the same cluster.

Mathematically, let $P = \{P^{(1)}, P^{(2)}, \dots, P^{(T)}\}$ be a set of T partitions, where each partition $P^{(t)}$ assigns each observation to a cluster. For each pair (i, j) , let indicator function $I_{ij}^{(t)} = 1$ if observations i and j are in the same cluster in partition $P^{(t)}$, and 0 otherwise. The terms of the co-association matrix $C_i = C_j = C_{ij}$ are then:

$$C_{ij} = \frac{1}{T} \sum_{t=1}^T I_{ij}^{(t)} \quad (3.12)$$

This matrix can be interpreted as a data-driven similarity measure, where higher values indicate stronger evidence that two observations belong together based on the ensemble. It is typically symmetric and can be viewed as a dense similarity graph.

The co-association matrix provides a flexible and model-agnostic foundation for consensus clustering. It allows integrating results from heterogeneous clustering algorithms and parameterizations, capturing the structural consensus across diverse perspectives. However, due to its density and pairwise nature, it requires careful handling for large datasets (e.g., thresholding or sparse approximations). Once constructed, clustering can be applied directly to the matrix using algorithms such as spectral clustering, hierarchical clustering, or density-based approaches, turning pairwise co-membership frequencies into meaningful cluster assignments.

3.5 Implementation

This section describes the intermediate results, modeling decisions and other issues encountered during feature extraction, ensemble generation, and consensus clustering.

3.5.1 Correspondence Analysis Calibration

Upon initial execution of the correspondence analysis using 2021 annual flows (n=1686), the data revealed distinct outliers and distortive subgroups, prompting targeted refinements to the sample. See Appendix B –C for more details.

After these adjustments to the sample, 1329 observations remain. The CoAn transform gave good results in that it is balanced with material vectors and facility type profiles apparently spread in almost all directions about the origin

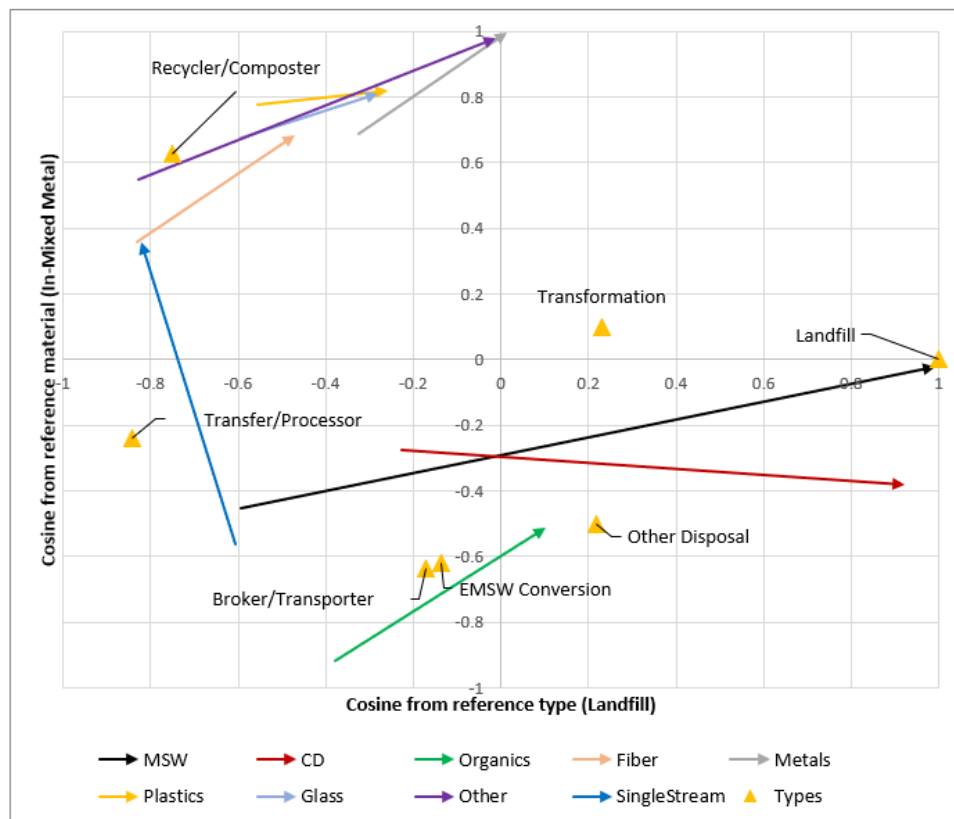


Figure 3.2 Biplot of Activity and Material Alignment in 2 Dimensions (Unit Circle)

Figure 3.2 shows this data further simplified into coarser categories and projected to a 2D biplot using two cosine reference vectors – one facility type (Landfill) and one material (In-Mixed Metal). Landfill is selected as the most distinctive facility type and In-Mixed Metal selected for maximum orthogonality to Landfill (Landfills do not take in mixed metal). This projection is mathematically suboptimal in terms of the distance representations but more clearly illustrates the basic trends of vector alignment. The arrows in Figure 3.2 each represent a material group. The start and end location of each arrow denotes the Out- and In- point, respectively.

Taking MSW as an example, the arrow's origin is in the lower left quadrant. Its proximity to the Transfer / Processor node reflects that most of the outflows of MSW originate with transfer stations. Similarly, its terminus, representing In-MSW is almost perfectly aligned with the landfill node. Compared to other materials, MSW, CD, and Organics are relatively aligned in terms of the directionality of the outflow-inflow relationship as well as the quadrants.

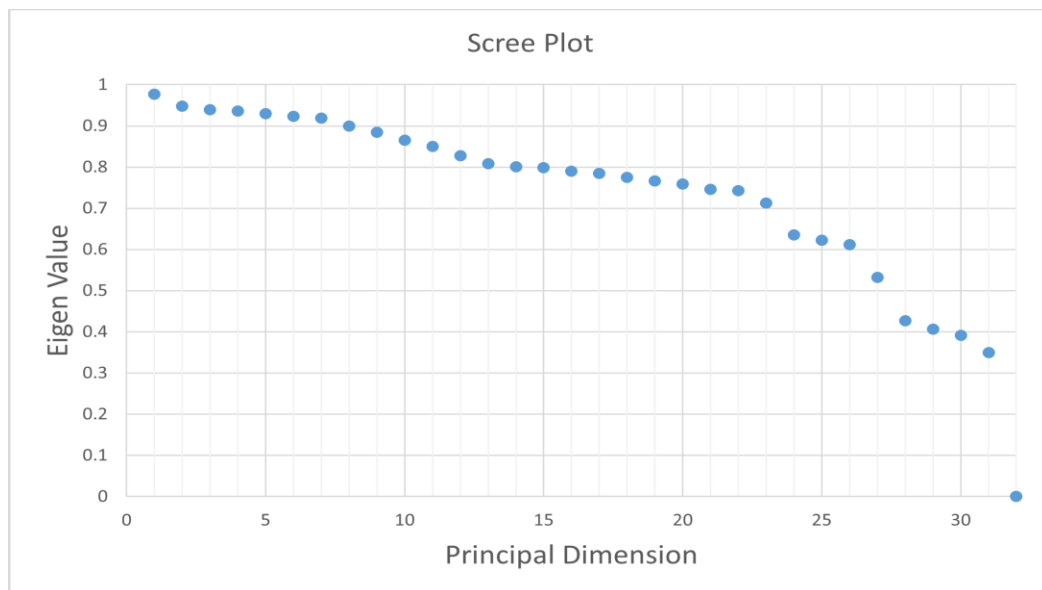


Figure 3.3 Scree Plot of Feature Dimension Eigen-values

Single stream, on the other hand, begins near the transfer/processor node but is clearly going in a different direction with its terminus in the Recycler/Composter quadrant. The rest of the recyclable materials (fiber, metal, plastic, glass) are all clustered in that same area, with outflow-inflow oriented approximately parallel in the same direction.

Organics is an interesting case because it is located very far from the Recycler/Composter node but retains the orientation of the recyclable materials. This speaks to one of the main goals of the reclassification, to separate Recyclers from Composters.

Examination of the eigen structure, represented in Figure 3.3 indicates a very gradual decrease in eigenvalues, with no single, unambiguous point at which to truncate the dimensionality. This is in contrast to textbook examples of scree plots because this involves 32 dimensions of high rank, a fairly substantial sample, and has already had its sample modified by the heuristics employed during CoAn calibration.

Cumulative inertia, defined as the proportion of total variance in the data explained by the leading principal components, provides one way to evaluate how many dimensions to retain. Because each eigenvalue represents the variance captured by a corresponding principal dimension, cumulative inertia is obtained by summing squared eigenvalues up to a given dimension and dividing by the total sum. Selecting a threshold for cumulative inertia is inherently an approximation, since there is no universally correct level of variance that guarantees adequate representation of the system. Instead, the threshold reflects a balance between parsimony and explanatory richness: higher thresholds preserve more information but at the cost of added complexity.

Careful examination of the cumulative inertia chart, represented in Figure 3.4 indicates a slight inflection point between the highest-eigenvalue principal dimensions (1 through 7) and mid-value dimensions (11 through 19). The choice to retain nine dimensions aligns with this inflection point while also satisfying an arbitrary but pragmatic 40% cumulative inertia benchmark. This compromise captures the essential structure of the system without overfitting to noise, providing sufficient richness to support multiple classification methods.

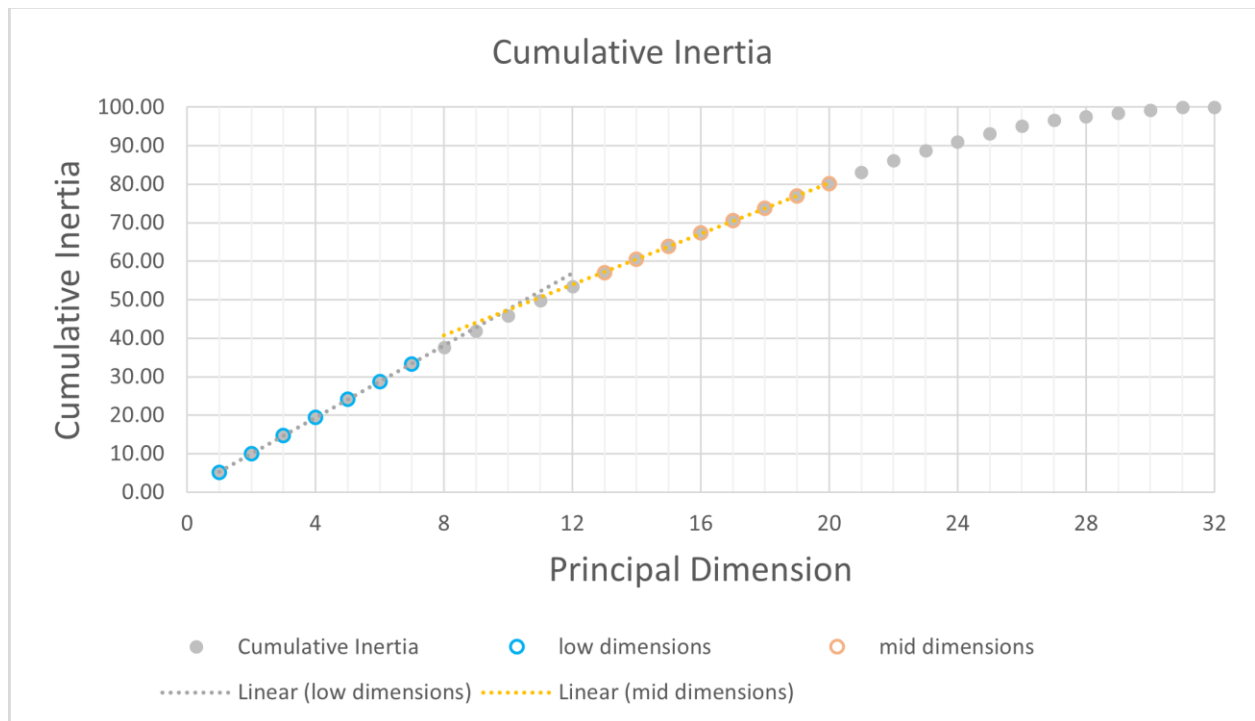


Figure 3.4 Cumulative Inertia Chart and selecting dimensionality reduction

3.5.2 Identifying Reference Vectors

Reference vectors used in the cosine similarity calculation are chosen to have the highest orthogonality between dimensions in the CoAn-transformed feature space. They are constrained to be observed facilities and are selected by evaluating a 2-dimensional cosine between a vector

of ones and each observation for every adjacent pair of retained dimensions. The minimum of the absolute value of the cosine measures is assigned to become the reference vector for that pair of dimensions. One final evaluation is performed to obtain the most orthogonal across all retained dimensions, simultaneously.

The resultant selections are listed in

Table 3.3. Each pair of dimensions under which the facility is deemed “most orthogonal” consistently have similar magnitude but opposite sign. The final All-to-All reference facility exhibits the widest spread of values. Interestingly, looking at the names of the facilities selected, this automatic selection process chooses very different facility types. This bodes well for its utility in discriminating the facility types. The only exception evidenced in the facility descriptions in

Table 3.3 appears to be the D7vD8 and D8vD9 vectors which both handle organics – they are the most similar of the references selected but do have differences in some of the dimensions as well. This suggests a potential refinement to the automatic selection procedure that chosen references could be compared to each other to ensure they are not too similar to each other. This could be enacted by some maximum threshold on the cosine or cosine-squared value between any pair of reference vectors. This would also prevent the same reference being selected twice, a more problematic possibility which did not occur in this dataset. If a suitably “different” reference for a pair of dimensions cannot be found, then refining the criteria to be most orthogonal across the three problematic dimensions is a minor refinement. Perhaps this is best practice as more dimensions are retained – after all, the problem did not seem to arise until higher dimensions – mathematical experimentation on a rich dataset with ground truth labels would be necessary to analyze this.

The final step of the feature extraction is executed by calculating the nine-dimensional cosine between each observation and the nine references selected. The BasisLogZ (log-transformed, z-scored facility size) is appended to the feature matrix to include scale information. The parameters of this transform for the facilities with valid patterns are $\mu_L = 3.87$ and $\sigma_L = 1.17$.

Table 3.3 Optimal feature reference vectors across 9 dimensions

Facility Name	D1	D2	D3	D4	D5	D6	D7	D8	D9
Rialto Bioenergy Facility; In-Vessel Digestion Facility Permit; RD12972	1.000	-0.132	-0.273	-0.008	-0.269	-0.131	0.500	0.547	-0.135
POLYWIZE; POLYWIZE - Transfer/Processor; RD13076	-0.132	1.000	-0.082	-0.094	-0.160	-0.262	-0.121	-0.087	-0.046
Weaverville Transfer Station; Medium Volume Transfer/Proc Fac; RD10054	-0.273	-0.082	1.000	0.122	0.438	-0.338	0.169	-0.389	-0.150
City of Long Beach - SERRF; Southeast Resource Recovery Facility; RD10079	-0.008	-0.094	0.122	1.000	-0.254	-0.067	-0.223	-0.132	-0.077
Florin Perkins Public Disposal Site; Florin Perkins Public Disposal Site; RD10302	-0.269	-0.160	0.438	-0.254	1.000	-0.211	0.013	-0.279	-0.024
Chico Scrap Metal, Inc. 20th; Chico Scrap Metal South; RD11163	-0.131	-0.262	-0.338	-0.067	-0.211	1.000	-0.373	-0.143	-0.098
Agromin OC, LLC - Oceanside; Agromin OC- Oceanside Green Materials; RD10267	0.500	-0.121	0.169	-0.223	0.013	-0.373	1.000	0.822	-0.161
Organic Depot, LLC; Organic Depot, LLC - Recycler/Composter; RD11310	0.547	-0.087	-0.389	-0.132	-0.279	-0.143	0.822	1.000	-0.030
Glenn County Landfill; Glenn County Landfill; RD10015	-0.135	-0.046	-0.150	-0.077	-0.024	-0.098	-0.161	-0.030	1.000

3.5.3 Ensemble Generation

The combined feature vector is then subjected to various clustering methods and parameterizations to obtain a collection of classifications. The label assignments within these runs can then be compared to one another analytically by creating the ensemble coassociation matrix which is itself subjected to clustering to form a consensus. Each ensemble consensus cluster can then be described and analyzed and the confidence of a given member's cluster assignment assessed directly by the coassociation matrix. This enables fine tuning of the final cluster scheme.

The goal is to generate multiple strong candidate clustering partitions by different methods in order to create a diverse ensemble for consensus clustering. The methods used are amenable to the cosine-transformed feature data and provide flexibility for multiple parameterizations. They are K-means, Hierarchical, and Mean Shift Clustering.

3.5.3.1 *K-Means Clustering*

K-Means is an unsupervised learning algorithm used to partition a set of observations into k distinct, non-overlapping clusters. It is particularly effective when the goal is to group data points based on similarity in a Euclidean or near-Euclidean feature space. For this study, K-Means is applied to feature vectors derived from correspondence analysis (CA) and cosine similarity, with the goal of uncovering latent operational types among waste-processing facilities.

In this analysis, K-Means is applied to cosine similarity features derived from CA-transformed material flow profiles of waste facilities and a transformed size basis measure. This method requires an a-priori decision on the number of clusters. Silhouette score and Inertia is utilized to select the “optimal” k for each experiment. The settings for python's sci-kit learn

method utilize random initialization of cluster centroids. The best of 50 randomly initialized runs on a single k -value is used to calculate inertia and silhouette score for the following screening graphs.

To select an appropriate number of clusters:

- K-Means is run for $k \in \{2, \dots, k_{max}\}$ for screening
- For each, inertia and silhouette score are computed.
- The results are visualized to identify an optimal trade-off between model simplicity (low k) and cluster quality (high silhouette).

This systematic approach ensures that the chosen clustering solution captures meaningful distinctions in facility behavior, while remaining interpretable and robust to overfitting.

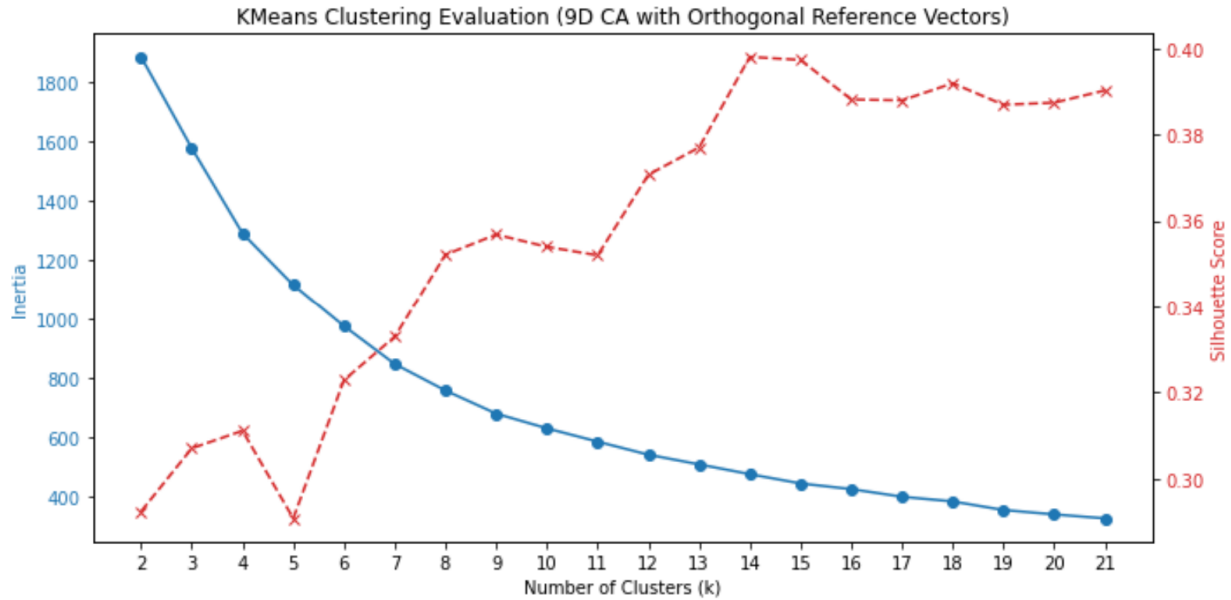


Figure 3.5 Inertia and Silhouette Score Charts for K-means screening runs

For this feature vector there is a very steady decrease in inertia with no discernable “elbow” in the Inertia curve on Figure 3.5 to recommend a value of k . And for silhouette score, there is very nearly a steady increase in silhouette all the way to $k=14$. A low, middle, and high value of k are chosen where the silhouette score peaks or plateaus, at $k=4$, $k=8$, and $k=14$.

3.5.3.2 Agglomerative Hierarchical Clustering

Hierarchical clustering is an agglomerative algorithm that builds a tree of clusters (dendrogram) based on pairwise similarities. In this approach, each facility effectively begins sorted into its own cluster. At each step, the two most similar clusters are merged based on a linkage criterion. The process continues until all facilities are merged into a single cluster, after which a threshold for the criterion or cut line is chosen, thus determining the number of mergings/clusters.

Hierarchical clustering is advantageous for its interpretability and for revealing nested cluster structures. However, it is more computationally intensive than k -means or mean shift for large datasets. The operation necessarily results in a tree of merges known as a dendrogram. Where to cut this tree can be determined post hoc and its visualization can support qualitative evaluation of cluster coherence.

The cut line, referred to as linkage distance, denotes the distance threshold for partitioning, below which observations are assigned the same cluster label. A higher value of linkage distance permits more clusters (further apart) to be merged together, generating fewer and larger clusters. Conversely, lowering linkage distance generates more clusters by permitting closer neighbors to be assigned to different clusters.

This method is implemented in Python using sklearn’s `AgglomerativeClustering` method for cluster output and `scipy.cluster.hierarchy` for dendrogram construction and visualization.

The pairwise similarity matrix is derived from cosine distances in the CA-transformed space. Clustering does not require a preset number of clusters; instead, the full dendrogram is produced once and can be cut at a desired depth based on visual inspection, silhouette, and/or linkage distance.

Figure 3.6 illustrates the first few branches of the merge tree. The y-axis shows the linkage distance and x-axis is a sorted representation of the cluster membership. An arbitrary linkage distance threshold of 2.0 is charted on the dendrogram, resulting in 5 clusters. Where the x-axis lists a number in parentheses, this corresponds to a count of membership further down in the tree, omitted for display. A number not in parentheses corresponds to an observation ID. For example, facility 1131 is the last merge enacted by the algorithm and stands alone as an apparent outlier, while the cluster represented by the left-most branch of the dendrogram has 568 members.

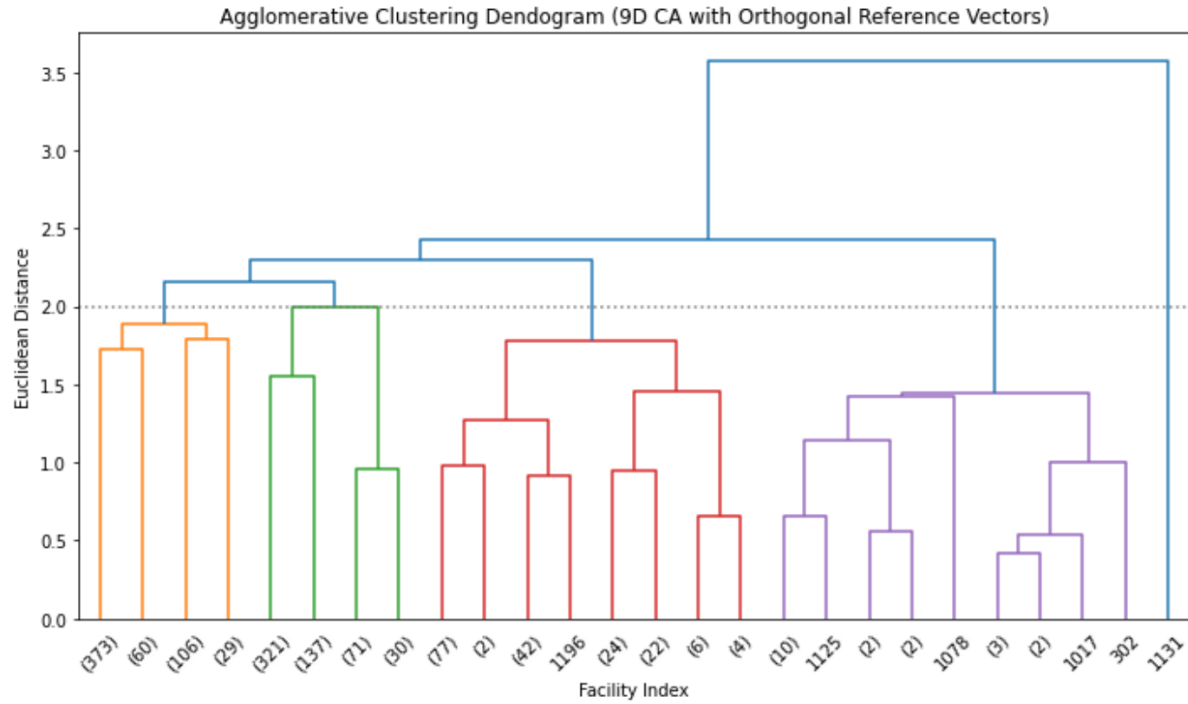


Figure 3.6 Dendrogram for Hierarchical Clustering

For the 9-dimensional feature vector, there is generally increasing silhouette score from $k=4$ to $k=12$ indicated in Figure 3.7. In this range, we select $k=7$ because it is the center of an apparent plateau in silhouette score from $k=6$ to $k=9$. Based on silhouette score alone, $k=8$ might be chosen but the very slight inflection of linkage distance at $k=8$ recommends $k=7$ (because the distance between seven and eight is greater than that between eight and nine). Examination of the dendrogram confirms that at $k=7$, one of the most populous groups is split up which is beneficial. The next time a single split effects many (>300) members occurs is at $k=11$, but the benefit to silhouette score from 11 to 12 is greater than that going from 7 to 8 so $k=12$ is chosen. At the high end of the screened k -values, there are a couple of clear peak/plateau points at $k=16$ and $k=19$. In summary, the parameterizations of $k=7$, $k=12$, $k=16$, and $k=19$ are selected for further analysis. In practice, the target number of clusters is implemented by setting the linkage distance.



Figure 3.7 Silhouette Score and Linkage distance Charts for Hierarchical screening runs

3.5.3.3 Mean Shift Clustering

To complement the partition-based (K-Means) and hierarchy-based (agglomerative) clustering strategies, Mean Shift clustering was applied to the cosine-transformed feature space derived from correspondence analysis. Mean Shift clustering is a density-based method which does not require the user to predefine the number of clusters. Instead, it automatically discovers clusters by identifying modes or local maxima in the data distribution. This makes Mean Shift particularly suitable for exploratory clustering tasks in high-dimensional feature spaces where the number of meaningful groups may be unknown or unstable across scales.

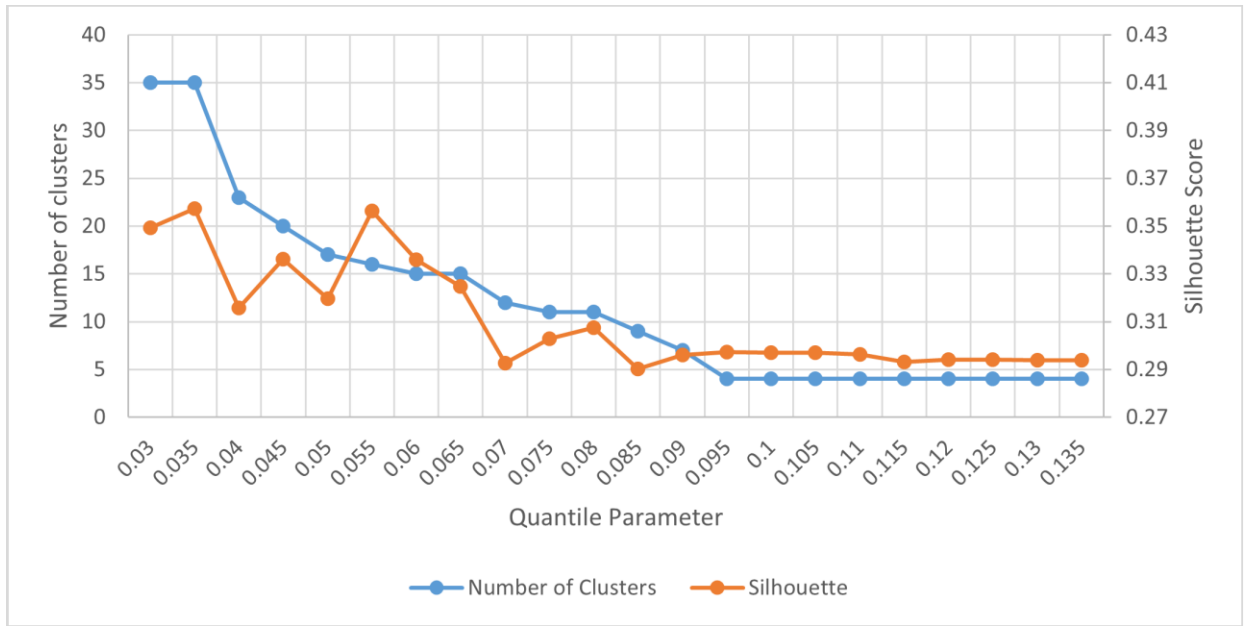


Figure 3.8 Screening of bandwidth quantile parameter for Mean Shift clustering

For the feature vector screening results shown in Figure 3.8, the following segments are proposed for further analysis due to their peaked silhouette score and apparent stability in k . For the very low quantile, high k range, silhouette seems unstable – a high value at 0.035 at first seems like peaking silhouette, but a large jump in number of clusters (from 23 to 35 in this case) naturally increases silhouette score. By contrast, at quantile 0.055, the number of clusters identified is fairly stable but a peak in silhouette stands out. This corresponds to 16 clusters and a bandwidth of 0.80. Similarly, a local maximum in silhouette occurs at quantile 0.08. Again, number of clusters seems stable at 11 but an increase in silhouette is registered. This corresponds to a bandwidth estimate of 0.97.

3.5.4 Ensemble Consensus Clustering

The ensemble utilized consists of a subset of all classifications run. These correspond to specifications which result in 11 to 16 clusters. This is slightly higher than the target number of facility classes. The runs included are:

- Hierarchical Clustering with cuts resulting in 12 and 16 clusters
- K-means Clustering with 14 clusters
- Mean Shift Clustering with bandwidth quantile settings resulting in 11 and 16 clusters

The small number of runs in the ensemble reduces the complexity of the coassociation matrix. Moreover, the similar range of clusters in each member run of the ensemble precludes the need for reweighting coassociations by cluster size. This allows slight differences in membership be due primarily to method and the within-run silhouette dynamics rather than the inverse relationship between cluster size and number of clusters.

The eigengap heuristic was used to select k . As seen in Figure 3.9 the difference between subsequent Eigen values falls off after $k=11$ and so this is selected for spectral clustering algorithm's internal processes (k for k -means algorithm and number of Eigen values retained).

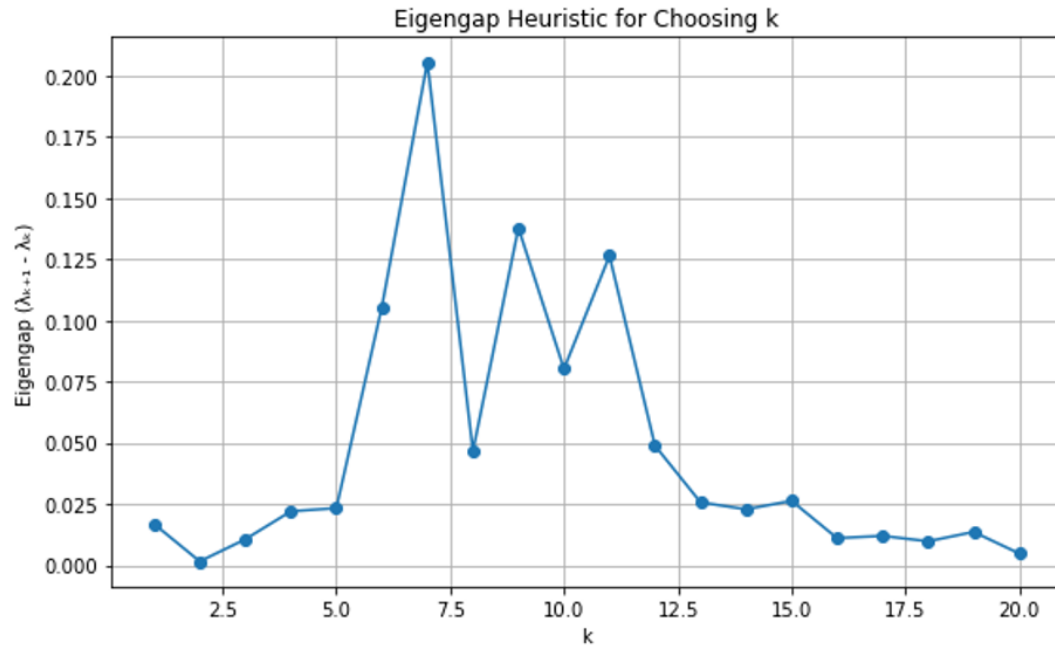


Figure 3.9 Eigengap chart for choosing k in spectral clustering

Executing the spectral clustering with $k=11$ produces a good result with respect to partitioning coassociations. Overall, the clustered coassociation matrix has strong concentration along the block diagonal as visualized in

Clustered items can be assessed a coassociation score between itself and its membership or individual belongingness in other clusters. Table 3.4 and Figure 3.10 illustrate the same data. Items are loosely ordered by their label assignment. The diagonal elements are strong indicating that each cluster has relatively strong agreement among members that those items belong in a group together.

Table 3.4 quantifies the average inter- and intra-cluster coassociation and makes it easier to see which nonzero inter-coassociations are actually significant. Note that coassociations less than 0.1 are indicated with red coloration, while those cells which are truly 0 are white.

Figure 3.10 shows a matrix heatmap of individual facilities, arranged by their cluster assignment. Strong coassociation appears as yellow coloration in the block-diagonal and weaker coassociations by greens and blues, particularly in the off-block-diagonal.

Table 3.4 Intra- and Inter-Cluster Coassociation

Cluster	0	1	2	3	4	5	6	7	8	9	10
0	0.7	0.0	0.0	0.0	0.0	0.0	0.1	0.0	0.0	0.0	0.0
1	0.0	0.9	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
2	0.0	0.0	0.9	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
3	0.0	0.0	0.0	1.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
4	0.0	0.0	0.0	0.0	0.7	0.0	0.0	0.0	0.1	0.0	0.0
5	0.0	0.0	0.0	0.0	0.0	0.8	0.1	0.0	0.0	0.0	0.0
6	0.1	0.0	0.0	0.0	0.0	0.1	0.8	0.0	0.0	0.0	0.0
7	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.9	0.0	0.0	0.0
8	0.0	0.0	0.0	0.0	0.1	0.0	0.0	0.0	0.8	0.0	0.0
9	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.6	0.4
10	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.4	0.9

Figure 3.10 indicates size of the consensus cluster via the dimensionality of the block diagonal square matrices. There are five clusters exhibiting very strong internal consensus (above 0.9) {labels: [1, 2, 3, 7, 10]}. The greatest inter-cluster coassociation occurs between 9 and 10, at a value of 0.4; those ultimately labelled 10 are solid (visible in the bold yellow square in the bottom right corner of Figure 3.10, but about 40% of the time 9's members are placed with 10's by one or more of the algorithms/parameterizations.

The moderately high intra- values (0.70 - 0.89) correspond with smaller inter-values (0.1) - these are 0x6, 5x6, and 4x8.- every cluster except 3 has some element of nonzero inter-

coassociation but this is not an issue. If a more diverse ensemble was selected, more elements in the matrix would naturally become nonzero.

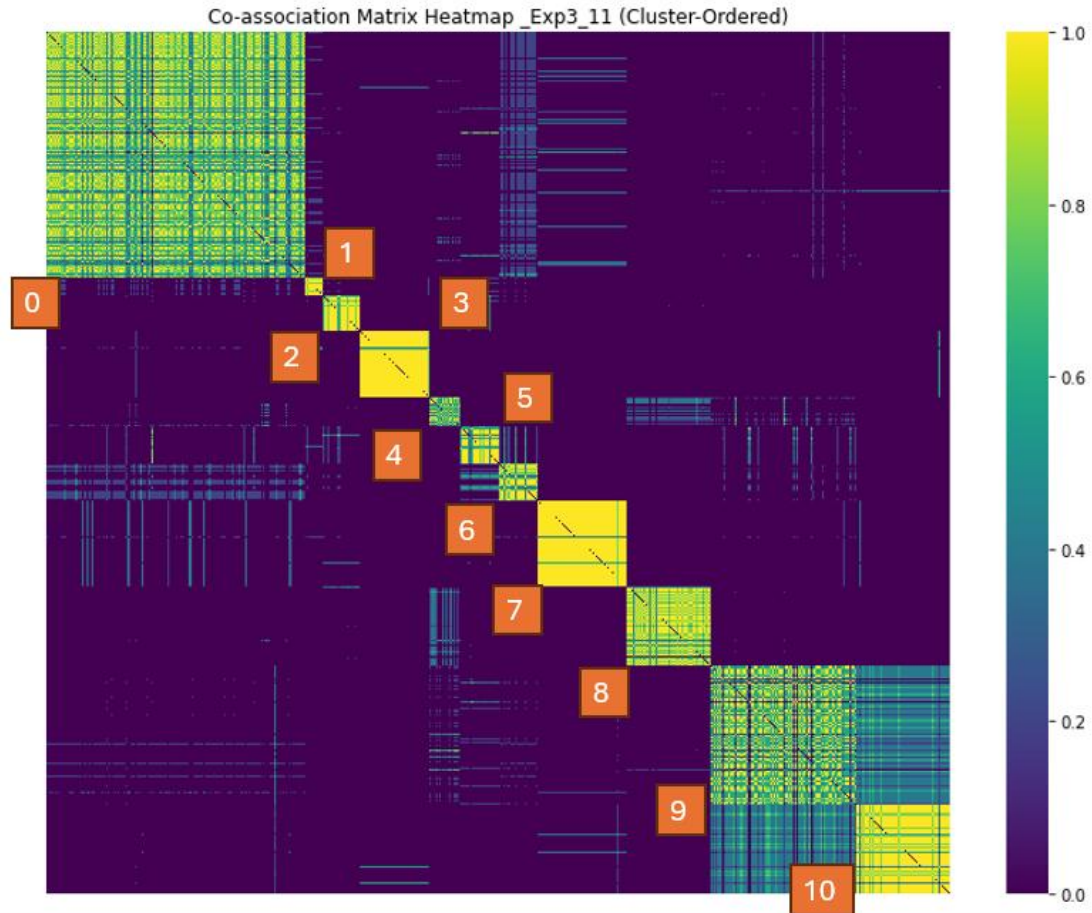


Figure 3.10 Heatmap of coassociation matrix, sorted by consensus cluster label

3.6 Cluster Descriptions

Next, a description of the clusters, outlining the defining features of each as well as the differences which discriminate them, is offered.

Examination of the forcing data, the pattern vectors themselves, is the most interpretable data.

An isometric log ratio (ILR) model is applied to calculate mean and 95% confidence intervals on the compositions (see

Appendix D –). The log transform on size basis is also inverted given the forward transform mean and variance parameters ($\mu_L = 3.87$ and $\sigma_L = 1.17$). These metrics are provided as quantitative descriptors of each cluster in Table 3.5.

Table 3.5 Composition and size profiles of clusters

Cluster	Membership	Estimated Mean Compositional Profile	Estimated Sizes (tons) [Lower CI95 bound, Mean , Upper CI95 bound]
0	379	9-24% Cardboard In 9-17% NonFerrous Out Recyclables In and Out	[132, 24552 , 165990]
1	28	99.80% MSW In	[28, 1756 , 10869]
2	55	~100% CD In	[30, 7658 , 52178]
3	101	99.97% MSW In	[12654, 505800 , 2933940]
4	45	85-98% Organics In 2-15% Organics Out	[16802, 131445 , 490317]
5	57	16-30% Metals In 9-22% Fiber In Misc. Materials In	[2, 56 , 304]
6	57	~96% Plastics In	[12, 2154 , 14538]
7	132	~100% CD Out	[4582, 86347 , 426902]
8	124	~100% Organics In	[31, 9705 , 66251]
9	213	53-60% Organics Out 38-43% MSW Out	[172, 25347 , 169542]
10	138	99.0% MSW Out 1.0% Organics Out	[27894, 245244 , 956271]

3.6.1 Landfills:

Landfills are represented in clusters 1 and 3. This is evidenced by the inflow composition of essentially only solid waste for disposal (MSW) and no outflows. The difference between the clusters comes primarily in the size basis. It is obvious that the members of cluster 1 are much

smaller in size whereas the landfills in 3 are much larger and vary much more in range. This is a product of the log transformation. These clusters have very high consensus.

Cluster 1 consists of very small sites characterized by primarily inflow of MSW. However, these are not all disposal sites – only 17 of 28 have active disposal permits. The remainder are transfer stations or other processors which are presumably too small to report their own outflows but do have other facilities reporting sending them material. The facilities in the latter case are marked for omission.

Members of cluster 3, on the other hand, are very consistent with SWIS labels tagging disposal operations. One broker/transporter is removed and two “transformation” sites are labelled as a new category (A third known transformation site is relabeled as well, from cluster 8). Ultimately, the two landfill categories, as they stand after the final cleaning operations mentioned above, are combined.

3.6.2 Transfer Stations:

Transfer stations handling MSW and other materials are labeled as cluster 9 and 10. Cluster 10 are larger facilities, more focused on MSW outflows while cluster 9 has an almost even balance of MSW and Organics outflows on average.

Analysis of their coassociation tendency as indicated by Table 3.4 and Figure 3.10 is informative. The weakest internal consensus in cluster 9 naturally corresponds to the highest inter-cluster consensus. Due to the functional similarity of Clusters 9 and 10, it makes sense that there is a larger area of overlap in the discrimination of these aspects across different methods. However, the mixing of associations between 9 and 10 is not symmetrical. That is, there is less doubt about those ultimately assigned cluster 10 than those assigned cluster 9. Almost 30% (59

facilities) of cluster 9 facilities have significant coassociations – this appears limited primarily to members coassociated with cluster 10, but there are non-zero associations to members of clusters 4, 5, and 6 as well. When members of mixed associations between the two transfer station categories are separated from pure, the mixed associations exhibit higher organics outflows (up to almost 87% for cluster 9 and 7% for cluster 10). Thus, coassociation with 4 is likely explained by the compositional similarity with composters (high organics) while coassociations with 5 and 6 are likely on the basis of size (very small). This is confirmed by cross-checking baseline RDRS activity category – some 54 facilities assigned to cluster 9 carry the Recycler/Composter activity category from the raw RDRS data. These have higher MSW outflows than those assigned composter categories by the ensemble clustering and very high organics compared to other members of cluster 9 and 10. These are presumably processing contaminated green waste streams and are reclassified as a third class of composters (labelled as ‘shifted’).

3.6.3 Construction Debris Inert Transfer and Disposal:

Construction and Debris (CD), also known as inerts, are handled by clusters 2 and 7. These clusters have very high consensus. Cluster 2 consists of exclusively CD inflows while cluster 7 consists of exclusively CD outflows. This indicates that cluster 2 consists of processors and end users (or more generally ‘sinks’) of CD while cluster 7 represents transfer stations specializing in CD. This is not to say that they flow exclusively from one to the other. The size distribution and count of membership (cluster 2 has fewer than half as many members and much less average throughput) indicate that only a small fraction of CD outflows from cluster 7 correspond with inflows to cluster 2.

Membership in cluster 2 is validated by pulling active SWIS permits for inert waste, processing or disposal. It seems that a number of facilities permitted to process CDI do not

actually do so regularly. Perhaps, they look for high quality supply (rich in metals or cleaner, more easily marketable building products). Many of those which do utilize their CDI disposal permits are landfills and are tagged as such (cluster 1 or 3). Many of those which utilize CDI processing permits are transfer stations and are tagged as such (cluster 7 if focused on CD, or clusters 9 or 10). The remainder comprise cluster 2 or are not reporting in the model year.

3.6.4 Composters:

Composters, characterized by handling of organics aka putrescibles, comprise clusters 4 and 8. Both clusters have primarily organics inflow but cluster 4 does exhibit some significant outflows of organics as well. Cluster 4 consists of larger facilities. These two clusters have significant inter-cluster coassociations. This appears mostly due to differing partitions on size across runs.

A subset of composters are biomass or anaerobic digestion facilities which perform special processes to stabilize waste and produce biogas. There are not many of these facilities in California, and when they are collocated with other activities, they are not typically separable on a process level from total reported flows in RDRS. SWIS was leveraged to identify these permitted digestion activities and they are tagged as type AD.

3.6.5 Processors and Recyclers:

Processors and Recyclers handling numerous materials are represented in clusters 0, 5, and 6. These are discussed together because they have high inter-cluster coassociations and similar materials in play, yet there are key differences which reveal their function in the broader waste system.

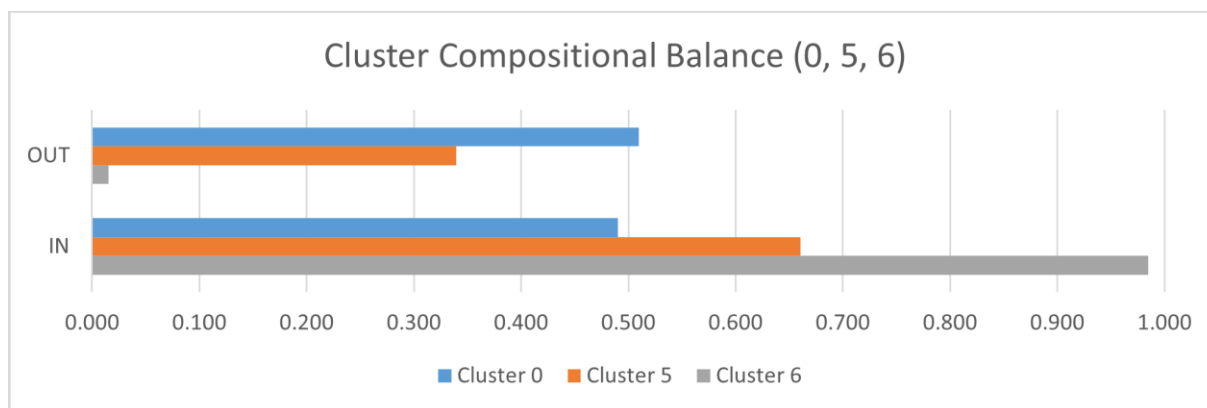


Figure 3.11 Gross inflow and outflow balances for processors/recyclers

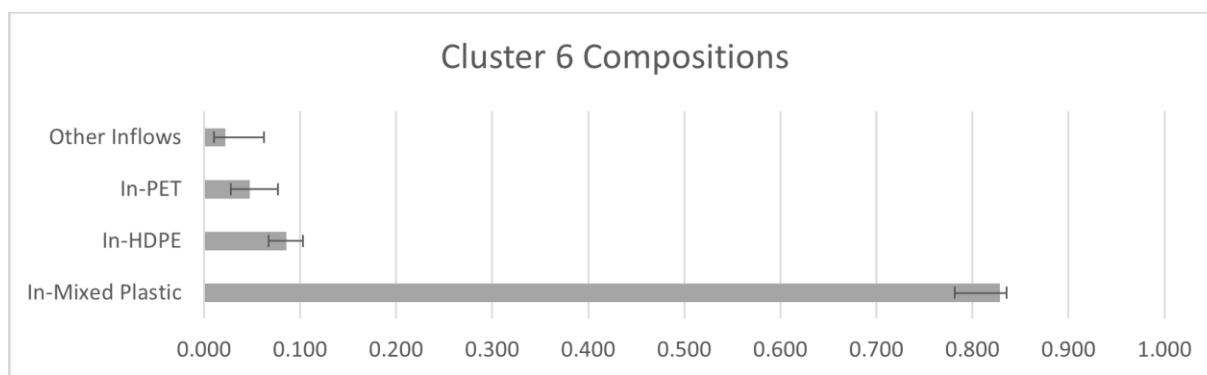


Figure 3.12 Compositional profile of Cluster 6

Cluster 6 has several features distinguishing it from the other two. First, the facility members are entirely sinks. In fact, the very small amount of outflows apparent in Figure 3.11 are just placeholders required to avoid $\ln 0$ in the ILR model. This means they do not report outflows but rather other facilities report sending them material. This material is primarily plastics (mixed, PET, and HDPE). It is important to note for these specialized facilities that mixed plastic can include either mixed streams or separated streams which are not PET nor HDPE (PP or LDPE, for example). In this context, it should be considered miscellaneous plastic and not necessarily mixed plastic – much of the total inflows to cluster 6 members are

mixed/misc, as indicated in Figure 3.12. Ultimately cluster 6 is best described as domestic endpoints of plastic flows and are labelled “Recycler (plastics)”.

Cluster 0 and 5 are very similar to one another but several distinguishing details can be found upon close inspection. First of all, cluster 0 is the single largest cluster with 379 members. By comparison, cluster 5 is one of the smallest clusters at 50 members. Operationally, their main difference is in terms of size. Cluster 0 consists of moderately sized facilities with a mean total basis of 24,552 tons per facility per year. However, the confidence interval on this ranges from 132 to 166,000 tons which is quite wide. Cluster 5, however, consists of the smallest of the small with a confidence interval from 2 to 304 tons and a mean of just 56 tons. When considering their compositions, it seems that cluster 5 takes in a greater variety of material types, with some construction inerts, putrescibles, single stream, and other materials entering. It has minimal outflows – the bars reported in the chart are almost all dictated by the zero replacement strategy employed to avoid $\ln(0)$ (Aitchison, 1994). Those with legitimate outflows have a visible upper confidence interval in Figure 3.13 (MSW, CD, Residuals, Mixed Plastic, Glass). There is a clear emphasis on fiber and metals for both clusters but glass and plastic inflows are represented as well.

Interestingly, cluster 0 and cluster 5 both have mixed coassociations with cluster 6 but not really with one another. These each have about the same level of intra- and inter-cluster coassociations (~ 0.75 intra- and ~ 0.12 inter-). The main reason 0 and 5 do not coassociate is because cluster 5 has a uniform distribution of outflows (very close or equal to 0) and thus is well-discriminated by the CA-cosine transform from cluster 0 which are clearly larger intermediate facilities. Examination of individual observations’ coassociations suggests that cluster 5 members coassociate with cluster 6 members on the basis of size regardless of

composition. The possibility that the smallest of cluster 0 members may also associate with cluster 6 due to size was explored, but there is little evidence of that when looking at aggregations of members with mixed coassociations, defined as $C < 0.6$. Furthermore, if this was the case, coassociation between 0 and 5 should arise. It is more likely that the lack of metal inflows (and therefore higher plastic contributions) and lack of outflows in general leads to some coassociation between 0 and 6 rather than a size effect.

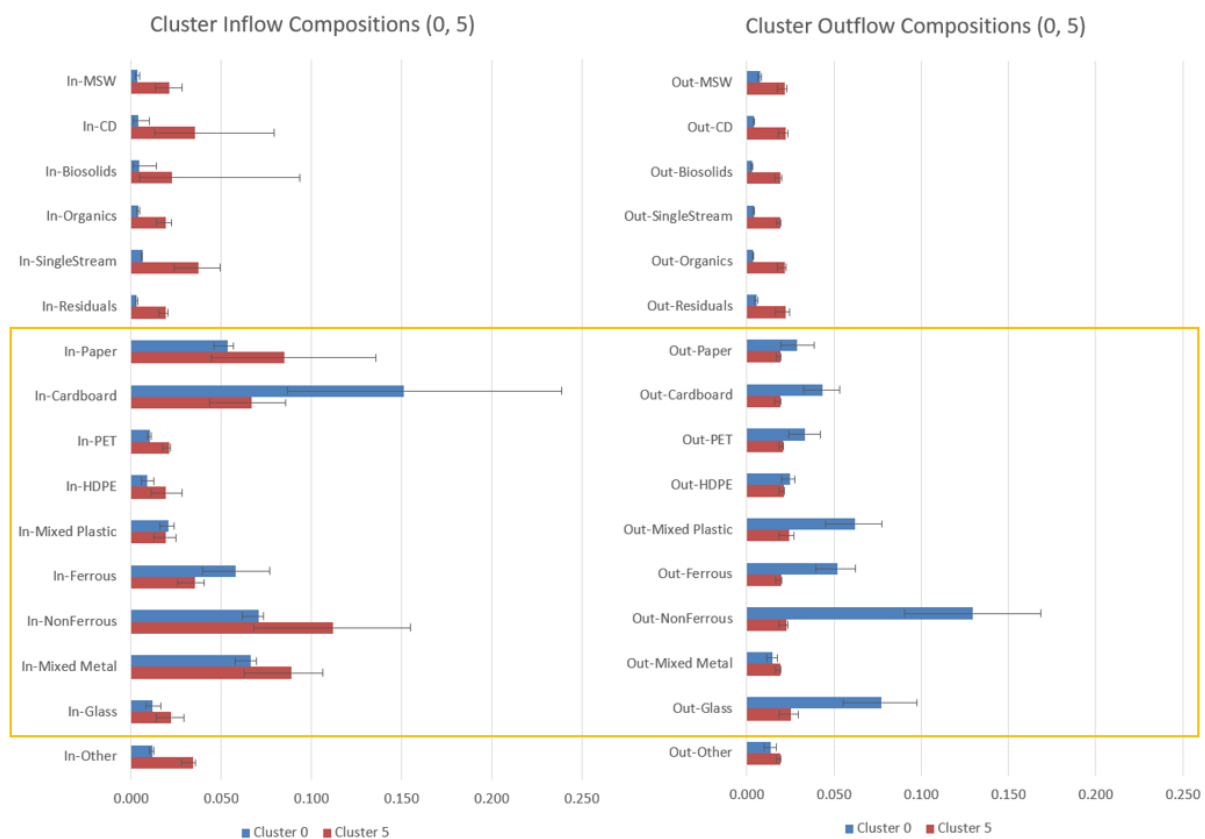


Figure 3.13 Compositional profiles of clusters 0 and 5

The large size of cluster 0, and its significant mixed coassociation patterns, suggest that it could be divided post hoc. Close analysis of the mixed coassociation members indicates that the mixed (lower) associations within the group are larger facilities focusing on paper and cardboard

while the higher coassociated facilities are smaller, focused on metals. In fact, these map to and fully contain two adjacent branches of a higher-k hierarchical run not included in the coassociation ensemble. That is, when a higher number of clusters is permitted, hierarchical clustering does split this group precisely along these lines. For now, this post hoc division of cluster 0 is applied for statistical analysis, primarily to break up the largest cluster. Thus, the cluster 0 is split into “Processors (small, metal)” and “Processors (medium, paper)”.

Cluster 5 consists of extremely small processors of recyclable materials. Several apparent misclassifications are corrected but the cluster is otherwise intact, relabeled as “Processor (very small)”.

3.6.6 Other Categories

Several categories were assigned upon seeing the initial correspondence analysis transformation and thus omitted from the clustering. There are 114 wastewater treatment plants identified, which exhibit large outflows of the organics subcategory of biosolids. A category of 69 facilities called Bulky Other was also separated, distinguished by flows of material type Other. 3 known Transformation sites and 15 Anaerobic Digestion sites (identified by SWIS permits) were separated as well.

3.6.7 Brokers

In addition to the activity designation of Broker/Transporter in the raw RDRS database, an auxiliary database of geocodes was utilized to perform a manual check of broker locations. This was done because it became obvious that many of them were residential or commercial (not industrial) spaces and therefore would not actually have waste-handling on the sites’ reported address. For this reason, brokers are pulled into a separate set of categories.

Brokers with addresses which are either co-located with waste operations or are an industrial warehouse with truck bays are separated by whether or not they have outflows reported. Most do not – 105 facilities are tagged as “Broker Sink” while 15 are labeled “Broker Intermediate”.

Brokers with a non-industrial site are labeled as a separate designation from other brokers because they cannot represent a movement of the material to the address in question. These are referred to as “Artificial”. Similar to industrial brokers, artificial broker/transporters are further separated by whether they have outflows (174 of 194 have only inflows). Those with inflows and outflows could potentially be repaired, replacing the reported legs through the broker intermediate with a single direct movement between the incident facilities.

3.7 Outcomes

The final classification of facilities leveraging the ensemble clustering, types of activities registered in RDRS and SWIS, and manual checks of certain types (e.g. brokers) results in the final breakdown of facility classes, presented in Table 3.6. The inflows and outflows to member facilities are tabulated to indicate systematic patterns in reporting as well as the portion of the total system for which that class accounts.

Next, the alignment of material movements with stated cluster descriptions is assessed by analyzing the origin and destination facility types of material movements. This is a two-step process to generate a disposition and label for each Origin Type, Material, and Destination Type combination.

Table 3.6 Summary of flow contributions by classification

Category	Membership	inflows (tons)	inflows (%)	outflows (tons)	outflows (%)
Landfill	118	43,205,491	75.97%	781,181	1.39%
TS(large)	138	1,951,289	3.43%	31,108,511	55.48%
TS(small)	153	89,030	0.16%	1,479,025	2.64%
TS(inerts)	132	907,121	1.60%	9,039,552	16.12%
InertsSink	49	233,702	0.41%	5,510	0.01%
compost(small)	54	304,259	0.54%	48,625	0.09%
compost(large)	38	3,704,406	6.51%	1,996,765	3.56%
compost(shifted)	59	164,575	0.29%	1,905,749	3.40%
Proc(very small)	42	1,983	0.00%	94	0.00%
Proc(small,metal)	179	159,142	0.28%	175,499	0.31%
Proc(med,paper)	152	2,072,652	3.64%	3,957,937	7.06%
Recycler(plastics)	38	50,802	0.09%	0	0.00%
WWTP	114	77,158	0.14%	2,442,690	4.36%
Transformation	3	613,031	1.08%	177,189	0.32%
AD	15	113,605	0.20%	57,608	0.10%
BulkyOther	69	50,215	0.09%	6,228	0.01%
Aritificial_NoRepair	174	2,220,829	3.91%	0	0.00%
Aritificial_CouldRepair	20	450,159	0.79%	2,154,367	3.84%
BrokerSink	105	385,921	0.68%	9,896	0.02%
BrokerIntermediate	15	101,670	0.18%	723,240	1.29%
OMIT	17	12,712	0.02%	678	0.00%
Total	1684	56,869,753	100%	56,070,345	100%

MSW moving from Transfer Stations to Landfills, Organics moving to Composters, etc.). There are 862 observed combinations assessed. Approximately 88% of the combinations are acceptable, but this translates to 99.7% of the total tons and no less than 97.8% of any one material included in subsequent analysis. This structure is evidenced in an updated (post-classification) version of the Activity and Material Alignment Biplot, Figure 3.14. This shows how relabeling by the ensemble consensus produces tighter alignment between material movements and the stated descriptions.

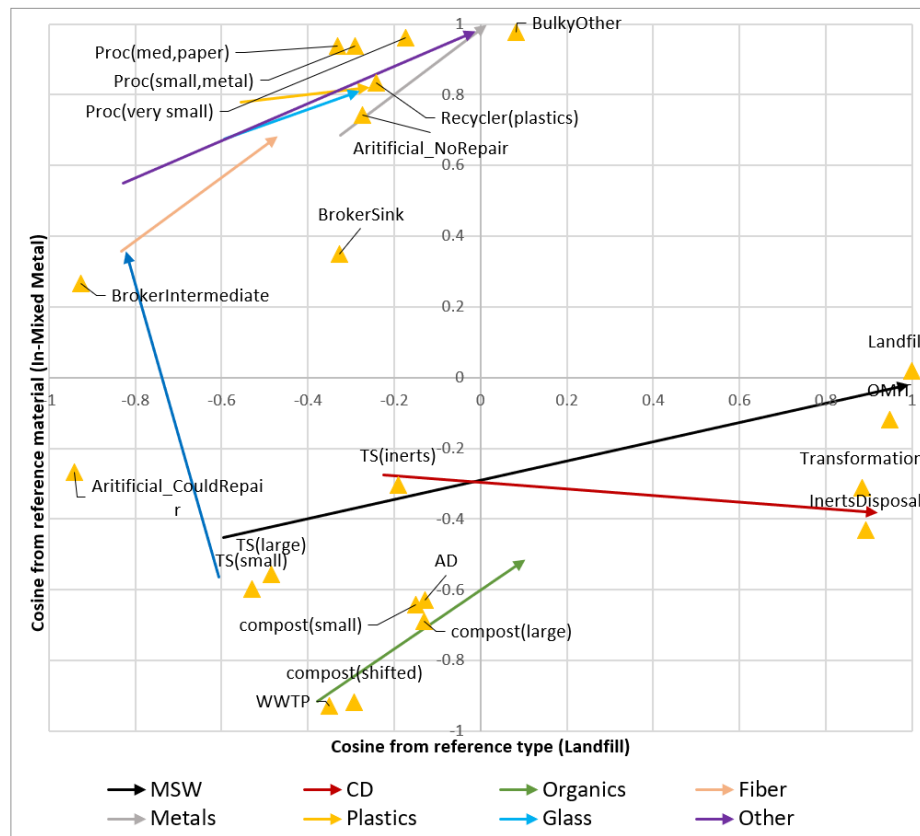


Figure 3.14 Biplot of Reclassified Activity and Material Alignment in 2 Dimensions (Unit Circle)

Second, every origin-material-destination combination is given a descriptive label which speaks to the nature of the facilities involved and/or material in question. They are: Disposal, Transfer, Composting, Diversion, and Broker Source or Broker Sink.

Taking the Origin Type of “Transfer Station (small)” as an example, less than 1% of flows by mass are marked anomalous and all dispositions are represented among flows deemed aligned. Any materials with destination “Inerts Disposal” or “Landfill” are labeled Disposal. The Transfer designation is made up of any material moving to any class of Transfer Station as well as any material that is *not* Organics moving to any class of Composter. Composting flows are more narrowly defined as Organics moving to any class of Composter, Anaerobic Digestion, or Wastewater Treatment facilities. Diversion is made up of any mixed or separated recyclable commodities moving to any class of Processor or Recycler. Broker [Source and] Sink includes any material moving [from or] to any type of Brokerage [origin or] destination, respectively. Flow labels are similarly defined for all origin types.

Finally, the ultimate goal of the classification exercise can be realized, which is to identify separate emissions model profiles for each facility class. These are shown in Table 3.7 and will be discussed in more detail in subsequent chapters. The transport between facilities is omitted for flows involving brokers and a few poor-fidelity sites tagged for omission. Operational emissions modeling (on site) is performed for the categories for which I have sufficient data, namely Landfills and Transfer Stations. This is further broken down by the source of emissions: heavy vehicle operations, landfill gas, electricity usage, and baling. The presence or absence of emission sources are stated for all classes as aspirational, even if I do not currently have data to estimate emissions in those activity spaces.

Table 3.7 Summary of emissions models by classification

Category	Transport Omitted	On-Ste Operations Omitted	Heavy Vehicle Activity	Electricity Auxiliary Equipment	Baling
Landfill	No	No	Yes	No	No
TS(large)	No	No	Yes	No	No
TS(small)	No	No	Yes	No	No
TS(inerts)	No	Yes	Yes	Yes	No
InertsSink	No	Yes	Yes	No	No
compost(small)	No	Yes	Yes	No	No
compost(large)	No	Yes	Yes	No	No
compost(shifted)	No	Yes	Yes	No	No
Proc(very small)	No	No	Yes	No	Yes
Proc(small,metal)	No	No	Yes	Yes	Yes
Proc(med,paper)	No	No	Yes	Yes	Yes
Recycler(plastics)	No	Yes	No	Yes	No
WWTP	No	Yes	Yes	Yes	No
Transformation	No	Yes	Yes	Yes	No
AD	No	Yes	Yes	Yes	No
BulkyOther	No	Yes	No	No	No
Aritificial NoRepair	Yes	Yes	No	No	No
Aritificial CouldRepair	Yes	Yes	No	No	No
BrokerSink	Yes	Yes	No	No	No
BrokerIntermediate	Yes	Yes	No	No	No
OMIT	Yes	Yes	No	No	No

3.8 Conclusions

This chapter has demonstrated how ensemble consensus clustering can be applied to California's waste facility data to overcome ambiguities in administrative classifications and to reveal operational archetypes grounded in material inflow and outflow patterns. By integrating correspondence analysis for feature extraction with consensus-based clustering, I produced facility groupings that are both interpretable and statistically stable.

The main contributions are threefold. First, I contextualized the facility classification problem within both waste management policy and clustering methodology, identifying gaps in

existing approaches. Second, I implemented a data-driven framework that adapts ensemble techniques to real-world infrastructure data, showing how consensus clusters capture differences in facility operations not apparent from regulatory labels. Third, I interpreted the resulting clusters as archetypes of California's integrated waste system, with implications for modeling emissions, evaluating diversion pathways, and informing policy.

The results highlight both strengths and limitations. On the one hand, the ensemble approach improved stability relative to single-method clustering and produced operationally meaningful groups. On the other hand, the approach remains constrained by data availability, assumptions about reporting completeness, and the challenge of aligning quantitative cluster descriptors with qualitative functional categories.

Future research could extend this work by (i) incorporating temporal dynamics to capture facility evolution or stability over time, (ii) integrating additional data sources such as facility permits, emissions inventories, or financial records, and (iii) exploring advanced consensus functions or optimization-based ensemble selection frameworks. These directions would enhance the robustness and applicability of facility classifications, helping bridge the gap between data-driven modeling and policy decision-making in waste system management.

Chapter 4 Quantifying the Carbon Footprint of California's Waste Management and Recycling Operations

4.1 Introduction

Building upon the research gaps established in Chapter 1, Chapter 4 develops the transportation and operational emissions framework for estimating the energy consumption and carbon footprint of household waste collection, recycling, and disposal, and presents statewide results. Although calibrated to California's regulatory and data environment, the framework is designed to be transferable to other jurisdictions and infrastructure systems.

LCA provides a quantitative basis for linking system activities to environmental impacts. In this study, GHG emissions are estimated as the product of activity levels (e.g., vehicle miles traveled, equipment utilization) and corresponding emission rates defined on a consistent unit basis. These estimates can be aggregated across processes to represent system-level impacts. However, the diversity of waste system operations means that no single dataset captures all relevant activities, necessitating an integrated modeling approach.

A central component of this framework is EMFAC, developed by the California Air Resources Board, which provides emission rates and inventories for on-road and off-road sources. EMFAC supplies emissions factors disaggregated by region, year, vehicle class, fuel type, and operating conditions, and serves here as the primary tool for translating modeled activity into GHG emissions.

The framework integrates three components of the waste system: (1) on-road collection, (2) post-collection transport, and (3) on-site processing. Collection emissions are obtained directly from the EMFAC Solid Waste Collection Vehicle (SWCV) inventory (Section 4.3.1).

Post-collection transport emissions are estimated by converting facility-to-facility flows into vehicle miles traveled (VMT) and applying EMFAC on-road emission rates. On-site processing emissions are estimated by linking material-specific processing pathways to equipment power demand and applying EMFAC off-road emission factors.

A key contribution of this work is the estimation of intermediate activity measures that are not directly observed. Specifically, routing models are used to estimate VMT for post-collection transport, and equipment utilization metrics (e.g., power expenditure per ton) are used to represent on-site processing intensity. These bridge quantities enable consistent application of emission factors across system components.

The analysis integrates facility-level and material flow data from the Recycling and Disposal Reporting System (RDRS) for calendar year 2021 with the facility classification framework developed in Chapter 3. This allows emissions to be estimated across heterogeneous facility types and material streams using consistent modeling assumptions.

The contribution of this chapter is the scale and comprehensiveness of my analysis, and the integration of waste flow data with modeling on-road and off-road waste processing and recycling operations. The study areas of published studies, even when they consider entire European countries, are smaller in scope than California, and the analyses of transportation and operations are typically based on simple multipliers such as average trip length and presumptive sequences of processing (Kirkeby et al., 2006; U.S. Environmental Protection Agency, 2016). Moreover, published empirical studies are usually limited to a small number of similar facilities and their focus is on operations, not policy. In terms of the system boundary I do not shy away from the complexity of the system represented in RDRS data nor the variety of activities undertaken throughout the state.

The research addresses two primary questions:

- (1) What are the total energy use and GHG emissions associated with the collection, transport, and processing of municipal solid waste and recyclables?
- (2) What is the relative contribution of transportation to the overall carbon footprint?

This work fills a key gap by leveraging comprehensive, system-scale empirical data rather than relying on limited facility-level observations or generalized assumptions. By integrating detailed material flow data with emissions modeling tools, the framework captures the operational complexity of the waste system and supports consistent, policy-relevant estimation of GHG emissions across interconnected processes.

The remainder of this chapter is organized as follows. Section 4.2 consists of a literature review spanning the many processes I researched. Section 4.3 identifies the emission quantification methods. EMFAC-derived and custom rate schedules are developed for multiple sources and species of emissions. Finally, in Section 4.4 Synthesis, the resulting emissions estimates from solid waste collection, post-collection transport, and on-site activity emissions, are combined. Results for logically partitioned subsectors, comparisons by scale of emissions, source of impacts are presented here. Contributions and opportunities for extensions are summarized offered.

4.2 Literature Review: Emissions Modeling Across Waste Management Systems

Life cycle assessment (LCA) has become the dominant framework for evaluating the environmental performance of municipal solid waste (MSW) systems. Comparative LCA studies across Europe and North America consistently demonstrate that system boundary definition,

substitution assumptions, transport distances, and temporal framing materially determine results (Buttol et al., 2007; Cherubini et al., 2008; Finnveden et al., 2005; Kirkeby et al., 2006). Carbon performance is therefore not an intrinsic property of a treatment technology; it emerges from how system components interact within defined analytical boundaries.

This section synthesizes the literature relevant to carbon accounting and infrastructure modeling in post-collection waste systems. It begins by examining life cycle assessment (LCA) frameworks and system boundary definitions, emphasizing how transport assumptions, substitution ratios, and methane management influence comparative outcomes. It then reviews carbon mechanisms by emissions pathway—transport, facility processes (recycling, biological treatment, construction and demolition), waste-to-energy integration, and landfill disposal—highlighting operational parameters that shape greenhouse gas performance. This section concludes by identifying limitations in existing approaches, particularly the reliance on stylized transport assumptions and fixed recovery efficiencies, and situates this chapter within these gaps.

4.2.1 System Boundaries and Carbon Accounting Structure

Across studies, recycling generally yields net climate benefits through displacement of virgin production (Finnveden et al., 2005; Krivtsov et al., 2004), while incineration with energy recovery typically outperforms landfill under fossil-based electricity systems (Buttol et al., 2007; Cherubini et al., 2008). However, rankings are sensitive to substitution context and transport assumptions. These findings motivate explicit decomposition of carbon impacts into transport emissions, facility process emissions, residual disposal emissions, and avoided burdens.

Most waste LCAs adopt a cradle-to-grave framing, tracking emissions from collection through treatment and final disposal (Finnveden et al., 2005; Wilson, 2002). Infrastructure construction

burdens are typically excluded, with operational emissions fuel combustion, electricity use, process emissions, landfill gas, and avoided burdens, dominating climate outcomes.

Comparative modeling frameworks such as WASTED and EASEWASTE explicitly integrate collection, transport, treatment, and avoided upstream production into a unified accounting structure (Kirkeby et al., 2006; Wilson, 2002). These tools emphasize that net greenhouse gas (GHG) impacts depend on:

- Waste composition (organic fraction, contamination, commodity materials)
- Energy recovery efficiency
- Substitution ratios for displaced electricity or virgin materials
- Methane capture rates in landfill
- Transport distance

In reality, carbon emissions from heavy-duty waste hauling are proportional to ton-miles transported, vehicle class and fuel type, load factor, and idle and queue time. Although many LCAs treat these factors generically, Brambilla Pisoni et al. (2009) show that collection vehicle emissions influence total impact profiles. Likewise, sensitivity analyses in Swedish and Danish LCAs indicate that increasing transport distances can diminish the relative advantage of recycling (Finnveden et al., 2005).

4.2.2 Transportation Emissions

Traditional LCA tools allow users to input transport distances but do not model routing decisions endogenously (Diaz and Warith, 2006; Kirkeby et al., 2006). Distances are scenario parameters rather than emergent outcomes of network configuration.

Transport is often parameterized in LCA models using assumed average distances (e.g., 15–40 km for treatment facilities and longer distances for recycling), yet sensitivity analyses show that increasing transport distances can materially alter comparative results (Finnveden et al., 2005).

Empirical and optimization studies provide more detailed insights. Brambilla Pisoni et al. (2009) demonstrate that collection and transport may represent a about one third of total system impacts (human health and resources).. Komilis (2008) shows that transfer stations become economically and operationally advantageous beyond threshold distances of approximately 50–65 km, influencing routing structure and haul intensity.

This framework extends existing approaches by incorporating empirically observed facility-to-facility flows and spatial routing to compute transport emissions explicitly. The following sections apply this methodology to each facility type, beginning with transfer stations and processors and proceeding through the remaining components of the post-collection waste management system.

4.2.3 Collection Activities

Collection has been studied a great deal as one of the earliest applications of operations research in waste management. Collection routing in a municipal context can be framed as an edge-covering problem and the routes are performed approximately once per week for as many “bins” as are being collected (Rhyner et al., 2017). In order to increase utilization of these vehicles, operators try to keep hauls short. The rolling stock to accomplish this typically run duty cycles of two or more truckloads (routes) per day. In terms of emissions, they typically have poor braking, idling, and speed characteristics, although a higher proportion of them run on natural gas fuel (Brambilla Pisoni et al., 2009; CARB, 2021).

4.2.4 Transfer Stations and Landfills

By contrast, interfacility transport typically benefits from load consolidation via transfer stations. Particularly with MSW, the use of transfer stations allows shorter hauls and higher utilization of collection vehicles as well as shorter queues at landfills. The economic favorability of this has been formally studied for some time (Komilis, 2008; Rhyner et al., 2017), evidenced in the fact that 60% of all (CalRecycle, 2023).

4.2.5 Composting Facilities

Organics processing facilities, including composting operations and other biological treatment technologies, are not explicitly modeled in the present analysis. The methodology developed in this dissertation relies on relating observable facility characteristics—primarily material throughput—to the equipment activity responsible for fuel consumption and emissions. For composting facilities, however, throughput is not a stable measure of operational demand because the material itself is continuously transformed throughout processing. Moisture loss, biological degradation, carbon conversion to gaseous products, and changes in bulk density all alter the mass and physical characteristics of the material between receipt and final product. Consequently, equipment activity cannot be reliably inferred from either incoming or outgoing tonnage alone.

Unlike transfer stations or landfills, where equipment activity is closely associated with material handling, composting operations depend on facility-specific practices such as windrow turning frequency, aeration methods, residence time, moisture management, screening operations, and product specifications. These operational variables are not captured within existing statewide reporting databases and would require a dedicated survey of facility equipment inventories and activity patterns to estimate with confidence. Because the objective of

this research is to develop an engineering-based inventory grounded in observable operational activity, inclusion of composting facilities would require assumptions substantially more uncertain than those employed elsewhere in the analysis. Their exclusion therefore reflects a limitation of available activity data rather than a lack of importance. Future work should extend the conditional simulation framework to organics processing through targeted facility surveys and should evaluate these systems concurrently with landfill gas and organics diversion scenarios, where important system-wide tradeoffs exist.

4.2.6 Construction and Demolition Waste Processors

C&D waste is dominated by mineral fractions and exhibits distinct carbon mechanisms relative to organics. Climate benefits arise primarily from displacement of clinker-intensive cement production and virgin aggregate extraction (Luo et al., 2024). However, realized carbon benefits depend on substitution ratios, market uptake, and transport intensity (Gálvez-Martos et al., 2018). In other words, diversion by weight does not necessarily bring proportional climate benefit, particularly when materials are downcycled. This may even be stipulated by law, at the expense of the contractor and the environment.

Construction and demolition infrastructure is likewise excluded from the present analysis because its transportation network and processing activities cannot be systematically reconstructed using available statewide datasets. Unlike municipal solid waste, which generally follows relatively well-defined pathways from collection to transfer and disposal, construction and demolition materials move through a decentralized and highly dynamic network of brokers, recyclers, aggregate producers, temporary crushing operations, and industrial end users. Materials such as concrete, asphalt, and inert fill may be processed at mobile facilities established for a single project before being relocated, while recovered commodities may

ultimately be transported to cement plants, road construction projects, private reuse sites, or other secondary markets that are not comprehensively documented.

As a result, neither the ultimate destination of these materials nor the equipment activity associated with their processing can be inferred from administrative reporting data alone. The principal challenge is therefore not emissions estimation but activity estimation: the locations, operating durations, equipment populations, and transportation pathways of many construction and demolition operations are effectively unobserved at the statewide scale. Developing an engineering-based inventory would require a dedicated characterization of these temporary facilities and secondary material markets. Consequently, construction and demolition infrastructure falls outside the scope of the present study, although the conditional simulation methodology developed herein provides a framework that could be extended to this sector as additional operational data become available.

4.2.7 Recycling and Material Recovery Facilities

Recycling's climate benefit arises primarily from avoided virgin material production. Early comparative LCAs consistently show that recycling paper, metals, and plastics reduces global warming potential relative to landfill and incineration under typical European conditions (Finnveden et al., 2005; Krivtsov et al., 2004).

However, facility-level performance depends on operational energy use, contamination rates, residual disposal, and market stability. Kerdlap et al. (2023) show that extrusion and processing equipment can account for a substantial share of operational electricity demand in recycling systems, and that centralized facilities achieve economies of scale relative to distributed systems. Vedantam et al. (2022) document a 23% increase in U.S. plastic landfilling following China's National Sword policy, illustrating how market instability alters realized

recovery outcomes. Turner and Kim (2024) note that U.S. recycling rates have stagnated near 35%, reflecting structural and economic constraints. These findings highlight that diversion rate alone is an insufficient proxy for climate performance; residual flows and substitution assumptions must be explicitly modeled.

4.2.8 Waste-to-Energy: Anaerobic digestion and Landfill Gas Systems

Waste-to-energy (WtE) encompasses a spectrum of technologies that convert waste-derived carbon into usable energy. While thermochemical pathways (mass-burn incineration, gasification, pyrolysis) and fuel production technologies exist, and I discuss some factors involved, these processes are less common in California. That is to say, MSW is not typically processed in this way in California to the degree it is in Europe or the East Coast.

Anaerobic digestion (AD) converts organic material into biogas through controlled microbial processes. Thus, unlike windrow composting, AD introduces an energy substitution pathway. When biogas displaces fossil-based electricity or natural gas, climate benefits may exceed those of aerobic composting under comparable transport assumptions (Shah et al., 2019).

However, net carbon performance depends on:

- Biogas yield,
- Methane leakage rates,
- Digestate management,
- Energy substitution assumptions,
- Transport intensity.

Reported biogas yields range from approximately 300–500 m³ per Mg of organic waste, with methane content typically between 65–75% (Nanda and Berruti, 2021; Rashad et al., 2026). Fugitive methane emissions are typically below 1 kg CH₄ per Mg in well-managed facilities

(Zeller et al., 2020), though performance varies with reactor design, feedstock composition, and operational loading rates.

Landfill gas (LFG) recovery represents a form of delayed biological WtE. Conventional landfills generate methane through anaerobic decomposition over decades. Bioreactor landfill configurations accelerate decomposition by moisture management and leachate recirculation, intentionally increasing gas generations to create adequate levels for combined energy projects. It is worth noting that projects intentionally enhancing methane generation are not common in California and the EU due to legislation and nowadays are more commonly pursued in developing countries as an opportunity for energy diversification. That being said, capturing LFG is an essential aspect of landfill life cycles, continuing for decades after closure and final capping (Rhyner et al., 2017).

Landfill gas typically contains approximately 50–60% methane, with methane possessing roughly 25 times the global warming potential of carbon dioxide over a 100-year horizon (IPCC, n.d.). Carbon performance of landfill-based WtE depends on:

- Gas collection efficiency,
- Gas treatment (flare or energy recovery),
- Gas generation capacity (composition of waste),
- Gas generation rate (temporal, environmental)

While energy recovery can partially offset emissions, uncaptured methane remains the dominant climate risk in landfill systems. Consequently, landfill-based WtE must be evaluated jointly with methane generation and capture modeling.

Thermochemical and fuel-production pathways (e.g., mass-burn incineration, gasification, refuse-derived fuel) reduce waste volume and generate energy, but their climate

performance depends strongly on substitution context and emissions control technologies. These pathways are referenced here for definitional completeness but are not the primary analytical focus of the present modeling framework.

4.2.9 Synthesis

Across facility classes, three themes emerge:

First, carbon performance is boundary-dependent and substitution-sensitive. Reported emissions vary substantially depending on whether studies consider direct operational emissions alone or include upstream, downstream, and avoided emissions associated with material recovery and energy substitution.

Second transport intensity and routing structure materially influence system outcomes, yet the magnitude of this influence remains poorly quantified. Most studies rely on simplified distance assumptions or aggregate transportation factors, leaving uncertainty regarding the contribution of empirically observed facility-to-facility movements and spatial routing to overall emissions.

Finally, diversion rate does not directly equate to climate performance. Higher diversion rates do not necessarily produce lower greenhouse gas emissions because environmental outcomes depend on the technologies employed, transportation requirements, material composition, and the counterfactual fate of diverted materials.

4.3 Methods

This chapter develops a unified methodological framework for quantifying energy use, fuel consumption, greenhouse gas (GHG) emissions, and selected criteria pollutants across the municipal waste and recycling system. The framework is structured around three primary system components: (i) collection activities; (ii) post-collection transportation; and (iii) on-site facility

operations. These components are linked through empirically observed material flows derived from the California Department of Resources Recycling and Recovery Recycling and Disposal Reporting System (RDRS), which provides the foundational data describing the movement of materials between facilities across the state.

The central objective is to translate these observed material flows and associated infrastructure operations into consistent, system-wide emissions estimates expressed in carbon dioxide equivalents (CO₂e), while also supporting estimation of criteria pollutants (e.g., NO_x, PM). To achieve this, the framework adopts a modular structure in which each facility type and process pathway is associated with a specific set of models and parameterizations reflecting underlying operational characteristics. These parameterizations may vary by facility classification, material type, geographic context, and, where applicable, temporal conditions for forecasting scenarios. In this way, the framework is designed to accommodate both observed system conditions and prospective analyses under alternative policy or operational assumptions.

Across all components, emissions estimation follows a consistent analytical form: activity levels are first quantified (e.g., vehicle miles traveled, equipment operating hours, material throughput), and then combined with characteristic emission rates appropriate to the technology, fuel, and operating context. This product is referred to as an inventory. While this general structure is common to all subcomponents, the data sources and methods used to estimate activity and emission rates differ substantially between collection systems, transportation networks, and facility operations. Accordingly, each component is developed and presented separately in terms of (i) input data, (ii) activity quantification, (iii) emission factor specification, and (iv) resulting emissions estimates, prior to integration and synthesis.

While carbon dioxide constitutes the overwhelming share of GHG emissions, methane (CH₄) and nitrous oxide (N₂O) are also formed in smaller quantities. These species are converted to CO₂-equivalent units using global warming potential (GWP) factors presented in Equation 4.1.

$$[\text{GHG kg CO}_2\text{eq}] = \mathbf{1}[\text{kg CO}_2] + \mathbf{28}[\text{kg CH}_4] + \mathbf{265}[\text{kg N}_2\text{O}] \quad (4.1)$$

The coefficients in Equation 4.1 denote 100-year GWP factors from IPCC Assessment Report 5 (AR5) and may be evaluated on any basis of consistent mass units (Myhre et al., 2013; Pachauri et al., 2015). Although the contributions of CH₄ and N₂O are minor relative to CO₂ for diesel combustion, their inclusion ensures methodological completeness and consistency with standard emissions accounting practice. I would even argue that this improves transferability toward any fleet-building efforts applying greater sensitivity of fuel economy to fuel types and the effects of age or control devices.

A defining principle of the framework is the explicit enforcement of additivity in emissions accounting. System boundaries are defined such that all modeled components are mutually exclusive with respect to the emissions they represent. This ensures that contributions from different processes and sectors can be aggregated without double counting. The material and facility classification schemas therefore serve not only descriptive purposes but also function as filtering criteria that partition the system into analytically independent units. Under this formulation, completeness of coverage is treated as secondary to clarity of attribution: omissions are preferable to overlap, as double counting introduces ambiguity in interpreting both baseline estimates and scenario-based changes. Where strict independence cannot be guaranteed, such dependencies are explicitly identified and their implications are discussed.

Overall, the methodology provides a scalable and internally consistent approach to estimating the environmental footprint of waste management systems at the state level. By coupling empirically observed material flows with process-specific emissions models, it enables order-of-magnitude estimates of system-wide impacts while preserving sufficient structure to support disaggregated analysis by facility type, material stream, and geographic region.

In Section 4.3.1 collection activities are represented by EMFAC's Solid Waste Collection Vehicle (SWCV) inventory, which explicitly encodes refuse-collection duty cycles, low-speed operation, and the unusually high penetration of alternative fuels.

Section 4.3.2 introduces post-collection transport between facilities. These trips involve a heterogeneous mix of vehicle classes and routing patterns that are not explicitly captured by any EMFAC inventory. In this module, EMFAC is used strictly as an on-road rate library, with emissions calculated by applying speed-, fuel-, and region-specific factors to vehicle-miles traveled estimated from RDRS flows and routing analysis.

Then, in Section 4.3.3, on-site material handling emission models at various facility types are identified. These include fuel and energy usage. Several categories utilize EMFAC off-road model rates for heavy equipment modeling and processing at transfer stations, MRFs, and landfills is dominated by heavy equipment that operates off-road and for which no waste-specific inventory exists. These emissions are therefore modeled using EMFAC's off-road rate framework, applied to independently estimated horsepower-hour activity. Relevant mathematical models are introduced or otherwise referenced if needed.

4.3.1 Solid Waste Collection

Collection represents the initial point of entry of materials into the waste management system and the first transportation process acting on generated waste. Operationally, collection services are typically organized into residential (single-family and multi-family), commercial, and self-haul (or contract-haul) sectors. These distinctions reflect differences in service arrangements, material composition, vehicle configurations, and duty cycles.

Collection Activity Estimation

Municipal collection is dominated by three primary configurations of solid waste collection vehicles (SWCVs): automated side-loaders serving single-family residential routes with standardized carts; front-loaders servicing commercial and multi-family dumpsters; and roll-off trucks designed to transport large containers, including compacting boxes used at institutional and commercial facilities. These vehicle classes collectively define the core municipal collection fleet responsible for servicing regular routes and satisfying near-complete spatial coverage of waste generation sources. In contrast, self-haul and construction and demolition (C&D) activities frequently utilize a broader set of general-purpose truck classes (e.g., dump trucks, flatbeds), which are not exclusively dedicated to municipal waste collection and are therefore not explicitly represented within the SWCV inventory. Given their more sporadic usage patterns, these activities are assumed to contribute a comparatively smaller share of total collection vehicle miles traveled (VMT).

Emissions from collection activities are taken directly from the California Air Resources Board EMFAC on-road model, specifically its Solid Waste Collection Vehicle (SWCV) On-Road inventory. With the settings described in Table 4.1, collection-related emissions are calculated from CARB's own internal vehicle population and activity model. The built-in

estimations of activity are formed by combining a DMV registration–based population with mileage-accrual rates (CalVIUS). Temporal allocation of VMT across speed and time-of-day bins using on-board diagnostics from what they call portable activity measurement systems (PAMS) produces a prior speed distribution. These permit development of forecasts tied to population and economic growth inputs (CA Department of Finance, 2024; CARB, 2021; U.S. Energy Information Administration (EIA), 2023).

Table 4.1 EMFAC On-road specification summary to retrieve T7 SWCV Inventory

Factor	Available Levels	Level Applied or Disaggregate
Version	EMFAC2017, EMFAC2021, EMFAC2025	EMFAC2021 v1.0.2
Output	Emissions, Emission Rates	Onroad Emissions (tons/year)
Region Type	Sub-area, County, MPO, Air District, Air Basin, Statewide	Sub-Area
Calendar Year	2000 through 2049	2021
Season	Annual, Summer, Winter	Annual
Vehicle Category (EMFAC202X)	CAIRP, OOS, Public, Concrete, Port, Solid Waste Collection, Dump, Tractor, Other by Class	T7 Solid Waste Collection
Model Year	1977 through 2051	Aggregate for Calendar Year
Speed	5-90mph (multiples of 5)	Aggregate
Fuel	Gasoline, Diesel, Electricity, Natural Gas, Plug-in Hybrid	Disaggregate

EMFAC encodes key operational characteristics of the SWCV fleet, including a low average speed (approximately 7.3 mph average speed) reflecting stop-and-go refuse collection duty cycle and a relatively high share of alternative fuels, particularly natural gas (CARB, 2021). For this study, the analysis is parameterized using calendar year 2021 and restricted to the SWCV truck class, which is assumed to comprehensively represent municipal collection

activities. For this component of the system, use of a pre-computed inventory is appropriate because CARB's internally modeled assumptions closely align with the physical realities of curbside collection. This also speaks to the long-term establishment of monitoring and planning municipal waste collection systems (i.e. AB939).

Within the overall methodology, collection is therefore treated as an exogenously specified emissions source, anchored in EMFAC's empirically calibrated inventory. This approach reflects both the relative maturity of collection emissions modeling and the absence of detailed route-level data in RDRS, allowing the analysis to focus modeling effort on post-collection transport and facility-level operations while maintaining a consistent and policy-relevant baseline for collection emissions.

These inventory-derived activity levels implicitly capture the intensive routing requirements of municipal collection systems, in which vehicles cover dense service networks multiple times per week, resulting in substantial cumulative VMT. EMFAC places the statewide VMT for solid waste collection in 2021 at a whopping 328.4 million miles.

GHG from Collection

Curbside GHG was estimated directly from EMFAC's published inventories for Solid Waste Collection Vehicles (SWCV) and serves as a baseline condition, reflecting CARB's modeling assumptions rather than observed or optimized routing. The total fuel consumption for 2021 is estimated at 63 million gallons of diesel and 33 million gallons of natural gas. Following guidance from the Intergovernmental Panel on Climate Change (Myhre et al., 2013), totals for carbon dioxide, methane, and nitrous oxide are converted to GHG emissions, expressed in

carbon dioxide equivalents (CO₂e). This translates to a total of about 1 million short tons or 960,500 metric tonnes of CO₂-equivalent emitted statewide in 2021.

Criteria Pollutants

In addition to greenhouse gases and fuel consumption, this work quantifies emissions of criteria air pollutants associated with transportation and facility operations. These pollutants differ fundamentally from greenhouse gases in both atmospheric behavior and regulatory significance. Whereas greenhouse gas emissions are aggregated according to radiative forcing potential and climate impacts, criteria pollutants are evaluated primarily with respect to localized air quality, atmospheric chemistry, and human exposure.

For interpretability, pollutant species are grouped according to their functional role in atmospheric processes. First, ozone precursor pollutants include nitrogen oxides (NO_x) and reactive organic gases (ROG), which participate in photochemical reactions responsible for the formation of ground-level ozone. Because ozone itself is a secondary pollutant formed through atmospheric transformation, direct ozone emissions are not estimated; rather, precursor emissions are reported as indicators of ozone formation potential. In heavy-duty transportation systems such as post-collection waste transfer, NO_x emissions are generally expected to dominate ozone precursor burdens due to the combustion characteristics of diesel engines, while ROG emissions tend to be comparatively smaller.

Second, particulate pollutants are quantified through estimates of both PM_{2.5} and PM₁₀ emissions. Particular attention is given to the distinction between exhaust and non-exhaust particulate matter. Exhaust PM reflects combustion-generated particulate emissions from engines, whereas non-exhaust PM originates from mechanical processes such as brake wear, tire

wear, and road surface abrasion or resuspension. In modern heavy-duty fleets, non-exhaust particulate sources may constitute an increasingly important fraction of total particulate burdens due to improvements in diesel after-treatment technologies.

Finally, additional combustion-related pollutants including carbon monoxide (CO) and sulfur oxides (SO_x) are estimated as indicators of combustion efficiency and fuel sulfur content, respectively. Although these pollutants are generally secondary considerations relative to greenhouse gases and NO_x in heavy-duty waste transport systems, they remain important for understanding localized air quality impacts.

4.3.2 On road post collection transport

Post-collection transport is modeled by translating reported material flows into network-resolved vehicle activity and then applying emissions factors to estimate greenhouse gas (GHG) impacts. This process links facility-level reporting data with a spatially explicit transportation network, enabling the estimation of vehicle miles traveled (VMT) and associated emissions at multiple geographic scales. The framework proceeds by constructing an origin–destination (OD) structure from reported flows, assigning routes between facilities using a routing engine, disaggregating those routes into spatially attributed segments, and finally applying emissions factors derived from EMFAC to compute GHG emissions. The result is a consistent representation of transportation activity that can be aggregated from the link level to subareas and statewide totals.

Activity System Representation: OD Construction and Routing

The first step in this process is to recast the reported material transfers from the Recycling and Disposal Reporting System (RDRS) into a origin–destination (OD) matrix, where each ordered

facility pair with nonzero flow defines a routing demand. Annual tonnage is converted to trip counts using an assumed payload of 18 tons per trip, representing typical heavy-duty tractor operations. This simplifying assumption reflects regulatory weight limits and industry practice while acknowledging the absence of publicly available facility-level payload distributions. The resulting OD structure defines both the spatial extent and magnitude of transport activity. To support this and subsequent operations, the RDRS data was embedded within a graph database representation, where facilities and flows (and soon, paths) are explicitly linked.

The routing procedures implemented in this work build directly on prior development led by Craig Rindt, whose OSRM-based pipeline provided the foundation for generating facility-to-facility paths. This framework was designed to support a broader class of system analyses, extending beyond the present application to include life-cycle scale elements such as exports, vehicle end-of-life processes, and additional material flows not represented in RDRS (e.g., material movements outside of CalRecycle’s purview). Within this dissertation, the pipeline is applied in a reduced form focused on domestic post-collection transport; however, the path generation component remains a critical linkage, enabling the translation of reported flows into network-resolved vehicle miles traveled (VMT). I contributed to data preparation and quality assurance, including furnishing geocoded facility inputs and conducting validation of routing outputs.

Accurate routing requires reliable spatial representation of facility locations. Facility records in RDRS are therefore geocoded and subjected to quality assurance, replacing reported addresses with verified latitude–longitude coordinates. Additional review of geospatial inputs and other time-intensive data gathering was provided by Dr. Saphores and members of the REMADE UCI research team. The resulting dataset, comprising approximately 1,600 facilities,

is spatially linked to jurisdictional layers including EMFAC subareas, counties, states, and national boundaries, enabling subsequent spatial attribution of transport activity.

OSRM computes shortest paths on a preprocessed road network derived from OpenStreetMap, where link costs are defined by roadway classification and associated nominal speeds, with turn penalties incorporated. Each route is returned as an ordered sequence of steps, which are treated as path segments. For each segment, OSRM provides distance and travel time, from which implied speeds are calculated. These travel times reflect free-flow conditions rather than observed congestion.

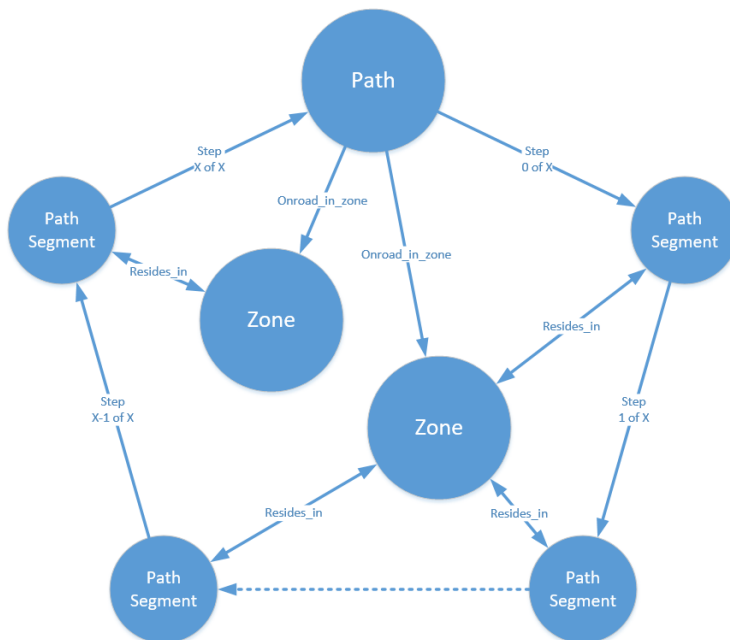


Figure 4.1 Graph Representation of Paths, Pathsegments, and TAZ (CoAbDis)

Path Segments are split (intersected) by the desired zonal layer – this includes geometry as well as distance and duration split proportionally assuming uniform speed. The Path Segments are

simultaneously assigned to the zone via a spatial join represented as a `resides_in` edge. Graph traversal from zones to incident Path Segments and associated step edges (which store `pathid`) and subsequent aggregation of distance and duration by path and zone lead to the creation of new edges (`onroad_in_zone`) directly from path nodes to zones as shown in Figure 4.1. This permits accurate estimation and aggregation of mileage and speed when the path involves segments in more than one CoAbDis zone.

Activity System Postprocessing: VMT Calculations

To support geographically resolved emissions accounting, each route is partitioned at subarea boundaries (CoAbDis). Path segment geometries are intersected with these boundaries to produce segments that reside entirely within a single zone, which I will refer to as links. This segmentation ensures consistent spatial attribution of VMT and compatibility with emissions factor lookup tables.

The segmented network is appended to the RDRS graph database, where PathSegment nodes are connected in sequence by directed edges. This structure preserves route topology while enabling efficient traversal and aggregation operations. Segment-level attributes include distance, duration, and implied speed, which serve as the basis for subsequent activity and emissions calculations.

I assume an average payload of 18 tons/trip to estimate the number of trips T_i from material flow weights. This value was chosen because it is close to the maximum gross weight allowed for non-drayage activities. Applying a different parameterization is as easy as rescaling. An estimate for a major subsector (such as transfer station to landfill movements) could be improved through surveys of payloads recorded exiting transfer stations and/or entering landfills.

This data should be readily accessible, likely digitized in most operations, but operators may not be amenable to data sharing.

Vehicle miles traveled (VMT) are computed at the segment level by multiplying segment distance d_i by the number of trips $T_{od} = Q_{od}/f_{payload}$ associated with the corresponding OD pair.

$$V_i = d_i T_{od} = d_i Q_{od} / f_{payload} \quad (4.2)$$

Number of trips is determined by an assumption of payload and applying a different parameterization is as easy as rescaling $T_{od} = Q_{od} / 18 * (\frac{18}{desired})$.

Emissions Rate Lookup: EMFAC On-Road

The next major step in the framework is to generate EMFAC on-road emission rates which scale on-road post-collection waste transportation activity (VMT) into GHG emissions. It is used strictly as a library of emission factors and not a pre-specified inventory. This is necessary because EMFAC does not define a single vehicle class exclusively representing post-collection waste transport (aside from Solid Waste Collection Vehicles, which apply to collection rather

Table 4.2 EMFAC On-road query specification summary to retrieve T7 Tractor Emission Rates

Factor	Available Levels	Level Applied or Disaggregate
Version	EMFAC2017, EMFAC2021, EMFAC2025	EMFAC2021 v1.0.2
Output	Emissions, Emission Rates	Onroad Emission Rates (mass/mile)
Region Type	Sub-area, County, MPO, Air District, Air Basin, Statewide	Sub-Area
Calendar Year	2000 through 2049	2021
Season	Annual, Summer, Winter	Annual
Vehicle Category	CAIRP, OOS, Public, Concrete, Port, Solid Waste Collection, Dump, Tractor, Other by Class	T7 Tractor
Model Year	1977 through 2051	Aggregate for Calendar Year
Speed	5-90mph (multiples of 5)	Disaggregate
Fuel	Gasoline, Diesel, Electricity, Natural Gas, Plug-in Hybrid	Disaggregate

than transload and long-haul movement) so a representative heavy-duty class is selected.

Emission rates for T7 tractor vehicles are extracted at the most granular available resolution in year, geographic subarea, speed bin, and fuel type; only age distributions and seasonal effects are

aggregated to annual averages. Table 4.2 indicates the appropriate EMFAC query criteria. This approach accommodates the heterogeneous mix of vehicle configurations observed in practice while maintaining consistency with EMFAC's rate structure.

At the link (path-segment) level, emissions are computed by mapping each segment's implied speed and geographic location to the appropriate EMFAC rate table. To apply these factors in my model framework, some aggregations are made. Speed bin and subarea will be retained as the two primary factors in the model and fuel type must be accounted for. Vehicle age is already aggregated by the EMFAC web export.

- Segment speeds s_i are discretized to the nearest upper 5-mph bin (e.g., a computed speed in (10, 15] mph maps to the 15-mph record/field), consistent with EMFAC's tabulated structure.
- Each segment is also associated with a specific subarea following the spatial join used to partition them by CoAbDis boundaries.
- Fuel-based differences in emissions rate (same speed bin, same zone) are handled by calculating VMT-weighted composite rates. See Appendix E for more information.

For each pollutant p , emissions on segment i $E_i^{(P)}$ is then calculated as the product of segment VMT (V_i), the corresponding EMFAC emission rate $F_p(t, s_i, z_i)$, and the weighting by fuel type $f_t(c, z_i)$, where s_i is the speed bin, z_i is the zone, t is the fuel type, and c denotes the vehicle class (see Appendix E). I do not extend this any further, but if more than one vehicle class is desired, their mix in/on a given zone/path can be modeled similarly to the fuel type weights.

$$E_i^{(P)} = \sum_t V_i * F_P(t, s_i, z_i) * f_t(c, z_i) \quad (4.3)$$

As an example, I will present the calculations for a link in Alameda County. In 2021, according to EMFAC's inventory, the share of total combustion miles associated with diesel and natural gas in that county was 90.8% and 9.2%, respectively. This weight matrix is used to create a composite emissions rate, the weighted sum of emissions rates of different fuel types represented in equation 4.3 as $E_i^{(P)}$.

For instance, consider a segment in Alameda County (id=54) which has a speed $s_i = 40$ *mph* and distance of 10.43 miles. Fuel fractions are calculated for Alameda and equation 4.3 can be rewritten:

$$E_i^{(NO_x)} = d_i T_i [F_{NO_x}(Diesel, s_i, z_i) * 0.908 + F_{NO_x}(NG, s_i, z_i) * 0.092]$$

The emission rates for the given fuels and parameter values are 'looked up' and the convex combination calculated. The NO_x production composite emissions rate of 2.00 g/mile.

Knowing this and the fact that the flow (id=5445) traveling on this path (id=96) is 117.4 tons of 'Other' material for disposal completes the calculation. Applying payload of 18 tons and equation 4.1, this translates to 6.5 trips, a VMT of 68.0 miles, and emissions of 154 g NO_x per year.

Other flows traveling the same path receive the same treatment, potentially with different parameters (e.g. vehicle type, payload). Also, segment-level totals for carbon dioxide, methane, and nitrous oxide are converted to GHG emissions (Myhre et al., 2013).

Link sequences are aggregated to within-county paths, at which point the Internal/External/Through disposition of the zonal vmt/trips is assigned. Care is taken not to double-count any emissions based on ambiguous joins.

This particular example raises an important point about the payload assumption and impact on trips. For this very small annual flow, is it reasonable that 6.5 trips were taken? Obviously, rounding up to 7 is appropriate in terms of accounting for that final half-full truck which still travels all the miles. But I suspect that this movement is (or could be) accomplished by more numerous, smaller payload trips – perhaps even a much smaller transport vehicle. It really depends on the material. For example, if this small quantity is disposed of weekly, then that is approximately 50 trips of a small 2 ton payload. This results in at least 7 times the VMT.

Statewide, the VMT, Trips, Fuel, and GHG are calculated for each county for a total of 142,900 metric tonnes CO₂eq in 2021.

4.3.3 On-site Activity Emissions

This is where various models are presented which contribute to on-site emissions. Similar to the on-road method, each is fundamentally the product of a species emission rate and estimated activity intensity. First, I outline the model for off-road heavy equipment emissions. I utilize EMFAC off-road to generate disaggregate emission rates and estimate activity to quantify waste-sector emissions.

Certain elements are absent for some facility types and others may have different values estimated for the same parameters.

4.3.3.1 EMFAC Off-road Heavy Vehicle Emission Rates

EMFAC Off-road module quantifies emissions from “yellow-iron” vehicles such as dump trucks and excavators as well as numerous other heavy equipment. The method is limited to compression-ignited combustion (diesel) and endeavors to simultaneously represent the impact of improved emissions standards implemented over the last twenty years as well as the detrimental effect of emission control deterioration across the long life that some vehicles are utilized.

Fuel use for off-road activities, upon which CO₂ generation and therefore GHG mostly depend, is modeled as a fixed rate. Diesel fuel expenditure is estimated using a uniform rate of 0.367 pounds per horsepower-hour (lb/hphr). This is converted to a mass-based GHG emission factor by first dividing by the fuel density (7.109 lb/gal) and then multiplying by the GWP-weighted sum of species emissions for diesel combustion (10.295 kg CO₂eq/gal diesel). This yields a GHG rate of 531.5 g CO₂eq per horsepower-hour of operation.

Emission rates for criteria pollutants, on the other hand, are more complicated to determine. While fuel usage (and therefore CO₂) is a fixed rate (per hphr) regardless of specific vehicle properties, criteria pollutants such as NOX and PM are modeled using deterioration rates. These are based on vehicle age and odometer as well as the specific emissions “tier” of the vehicle.

Emissions tiers correspond to increasing pollution control mandates. With the implementation of increased emissions standards approaching, vehicle manufacturing companies were allowed flexibility to engineer their own path to Tier 4 compliance (EPA, 2016). The levels are numbered from 0 (pre-regulation) through 4 (current regulation) and while they do not

strictly map to certain vehicle model years, for ease of discrete sampling, they are assumed to correspond to the following year mapping schema, presented in the leftmost two columns of data in Table 4.3. On the right side of this table, the EMFAC forecast-year vehicle population is used to calculate the fraction of vehicles in each range of ages or vehicle model years.

Table 4.3 Emissions Tier Base Year Population

Tier	From Year	Thru Year	BY2021 Population	Fraction
Tier 0	1900	1995	15867	0.0916
Tier 1	1996	2002	16976	0.0983
Tier 2, 3	2003	2010	37228	0.2150
Tier 4i	2011	2013	16268	0.0939
Tier 4_f	2014	2100	86835	0.5014

This means that some companies skipped certain tiers while others were allowed to step through intermediates to achieve Tier 4f. Tier 4i, for example, was allowed only if companies adopted early (skipping tier 3). Tiers 2 and 3 were combined due to similarity of rates and the overlap in years. I have assigned most of the tiers to appropriate but non-overlapping years so that I can easily sample the distribution of vehicle model years to replicate the tier distribution estimated for calendar year 2021 (or any desired application year).

For application year 2021, the calculations go as follows. The fractions of the population in each tier category are calculated. Each fraction is then allocated to representative vehicle model years to form a PMF. This implies that the age distribution of vehicle used for waste activities is the same as that for the overall construction population. Tier 0 is applied by a ramp from 0 in 1985 to the full mass fraction by 1995. Other tiers are assigned by a simple uniform distribution. Forecasts should probably adjust this aspect with respect to Tier 4f, perhaps by using a trapezoidal ramp or some other simple functional assumptions on vehicle age. In any

case, the PMF, shown in Figure 4.2 for Calendar Year 2021, can be sampled to simulate vehicle age.

Vehicle ages are important to model for criteria pollutants because each (NO_x , PM, THC, and CO) emission rate is determined by a baseline emission factor $F_0^{(P)}$ modified by a linear function of deterioration rate $F_D^{(P)}$, dependent on cumulative odometer (lifetime hours). This is assumed to be an average of 2000 hours per year over the life of the vehicle. Finally, the Fuel Correction Factor (FCF) is applied. The emission elements, $F_0^{(P)}$, $F_D^{(P)}$ are determined by vehicle

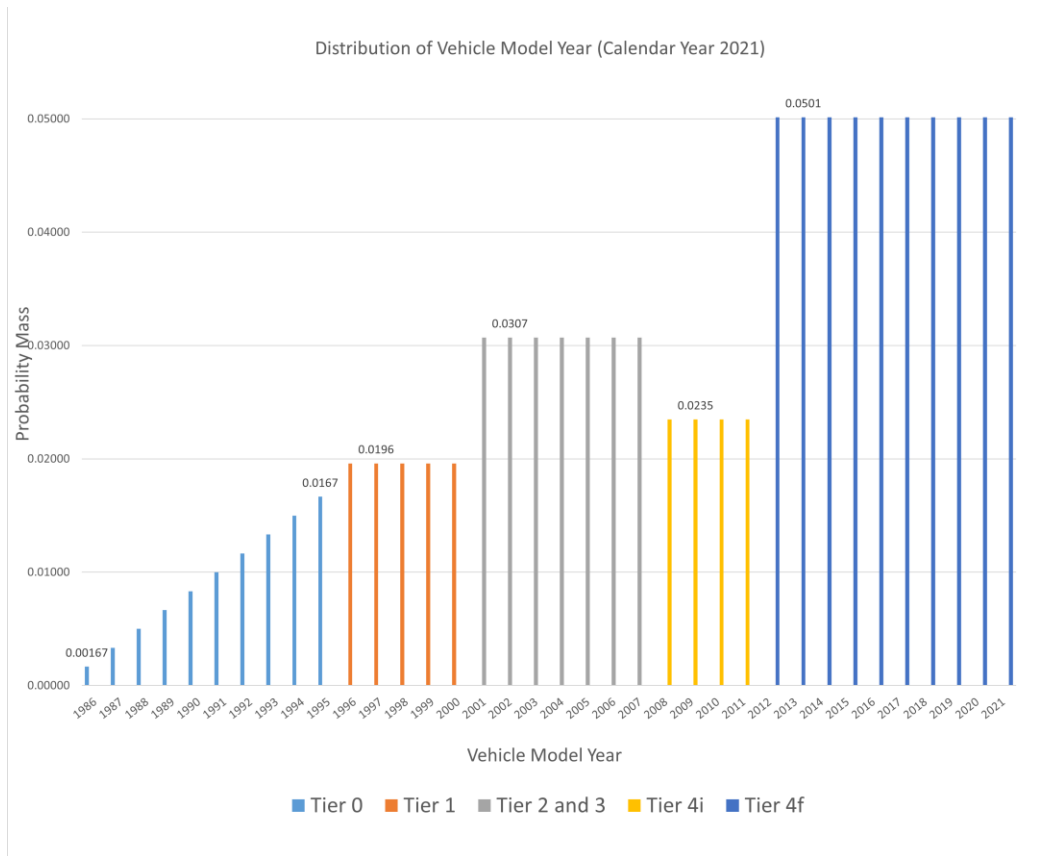


Figure 4.2 PMF of Heavy Vehicle Population Age (application year 2021)

model year (y) and vehicle horsepower, while the fuel factor, which accounts for changes in fuel characteristics over the years, depends on vehicle model year (y) and application calendar year (2021). Thus, the exact emission rate for criteria pollutant (P) for a given vehicle (v) are calculated by equation 4.5:

$$E_v^{(P)} = \left(F_0^{(P)} + F_D^{(P)}(2021 - y) * \frac{2000 \text{ hr}}{\text{year}} \right) * FCF \quad (4.4)$$

The resulting emission rates (g/hphr) by vehicle model year are shown in Figure 4.3.

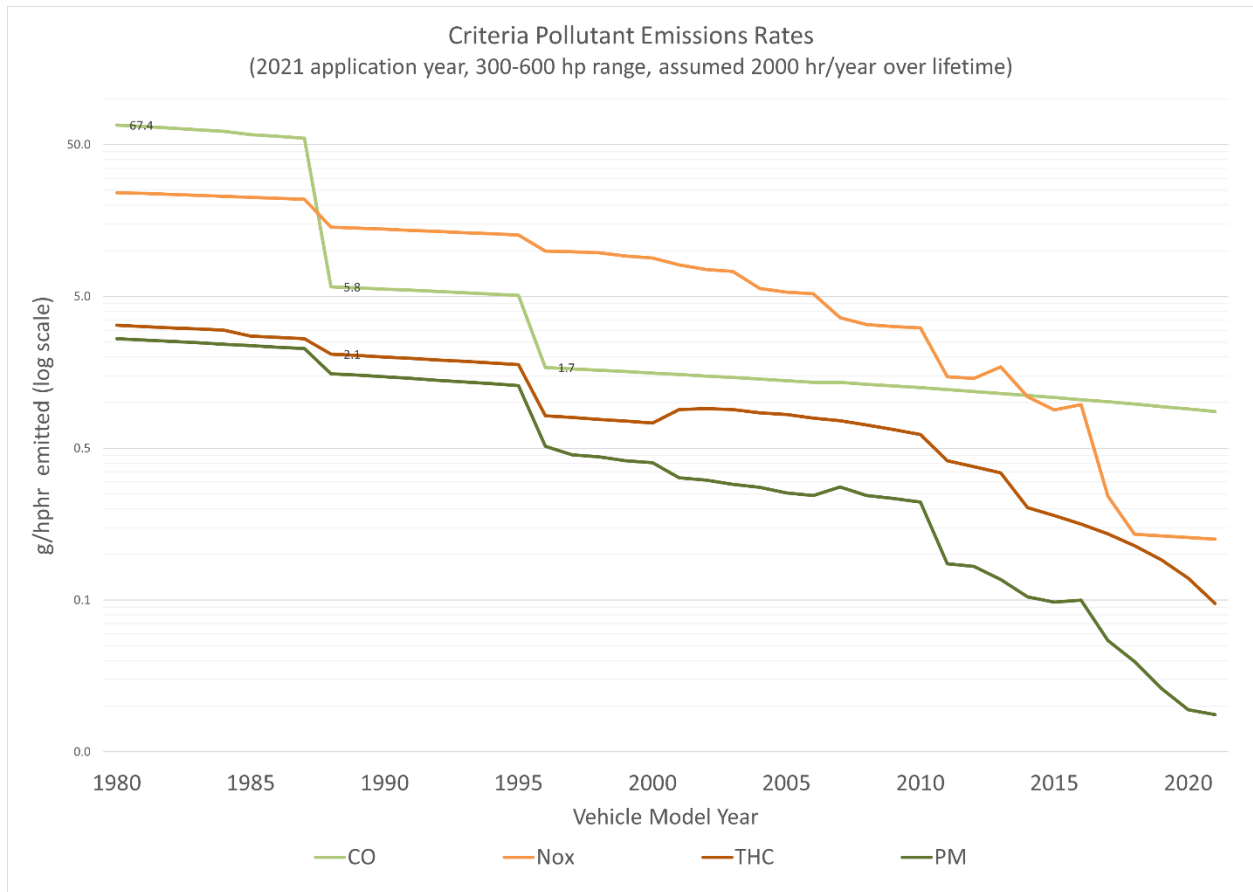


Figure 4.3 EMFAC Off-Road deterioration factors

With these established, a vehicle fleet (age, hp, type, and quantity) and its waste-related activities (hours operated) in 2021 could be used to generate an inventory.

4.3.3.2 Off-Road Activity Calibration and Fleet Simulation (HeaVe)

In this section, I estimate heavy off-road vehicle activities, in terms of horsepower-hours (hphr). The underlying assumption of this section is that, for a given activity (e.g. landfilling), the total power utilized is directly related to throughput quantity. Furthermore, there exists a stationary rate of power expenditure that is characteristic of certain activities and that different activities exhibit different rates.

To confirm these hypotheses and pursuant to activity estimation, I surveyed a number of facilities, measured their fleet configurations and vehicle usage, and estimated the characteristic power rates for their facility type via regression. Then, I applied those models to generate statewide activity estimates and ultimately, inventories of GHG emissions.

By measuring the total power expenditure and the total throughput over the course of a year, a point estimate for a facility-activity's average hphr (power) per ton can be established. To do this, I contacted organizations and operators to obtain information on the utilization of their heavy equipment. The survey included one transfer station, one material recovery facility, and eight landfills. More information about the survey and contacts are in Appendix B.1. I prioritized landfills for data collection because of the sheer volume of that subsystem as well as the variety of operations I observed during tours. Also, I suspected significant economies of scale so I wanted to survey the full range of landfill sizes, whereas I had only toured some of the largest landfills in Southern California.

The only aspect of equipment type that enters the activity calculation is the load factor. This number represents the intensity of power expenditure while operating. To be more precise, it may be thought of as a time-weighted average on the fraction of maximum power utilized during typical operation. So, during one hour of operation of a nominally 400hp machine, an assumed load factor of 0.4 implies a total power expenditure of 160 hphr. The load factors are all near 0.4 but the most relevant vehicles to waste operations are listed in Table 4.4 along with some exemplary vehicle models. The load factors are used to calculate the effective hphr from operations data provided by those surveyed.

Table 4.4 Heavy Equipment Load Factor

Equipment Type	Load Factor	Specific Models
Tractors/Loaders/Backhoes	0.3685	Cat440, Cat962
Skid Steer Loaders	0.3685	Cat242
Excavators	0.3819	Cat325
Off-Highway Trucks	0.3819	Cat745
Other Material Handling Equipment	0.3953	CatTH0842
Graders	0.4087	Cat120
Crawler Tractors	0.4288	CatD9
Off-Highway Tractors	0.4355	Cat836
Scrapers	0.4824	Cat623

Operational data on heavy vehicle equipment utilization was provided by several organizations. This data includes the equipment model and age and the annual hours of utilization.

The general procedure is to develop the average annual horsepower hours per ton of material handled multiplier (hphr/ton). This is done by summing the product of horsepower rating, load factor, and annual utilization (hours of operation) for each vehicle and then dividing

that total by the annual throughput for the facility. Mathematically, for a calibration facility of type f , the multiplier is calculated by Equation 4.6:

$$M_f = \frac{\sum_t^T \Pi_t}{Q_f} \quad (4.5)$$

where $\Pi_t = n_t p_t l_t u_t$

n_t = quantity of vehicle type t

p_t = rated horsepower of vehicle type t

l_t = average load factor of vehicle type t

u_t = annual hours of utilization of vehicle type t

Q_f = annual throughput of facility type f

Transfer Station Calibration

The transfer station calibration facility used is Irvine Transfer Station. That facility processes approximately 120,000 tons of waste each quarter and they do so, per their reported fleet data, with two 320hp wheel loaders and a 280hp excavator. Load factors for these types of vehicles, taken from Table 4.4, are 0.37 and 0.38, respectively (CARB, 2024). They operate six days per week (312 days/year) and 8 hours per day (Talavera, 2024). Dividing annual horsepower-hours by the annual throughput ($\sim 480\text{kT/y}$) provides average diesel engine energy demand per year. These can be evaluated per vehicle, per vehicle type (1.24 hphr/ton for loaders, 0.55 for excavators), or as a facility total (~ 1.8 hphr/ton). In the absence of a broader survey, assuming that this point estimate is stationary permits calculation of power demands for like facilities throughout the state.

MRF Calibration

The MRF calibration facility is WM's Orange/Anaheim MRF. We had received a tour and the supervisor (Ricardo) connected me with the maintenance supervisor who disclosed their heavy machinery usage statistics. MRFs have similar machinery requirements as transfer stations in terms of loaders and excavators but have less throughput naturally due to more complex operations. At this site, the fleet consists of two wheel loaders and an excavator, all with lower rated hp than Irvine Transfer Station but significantly more hours per vehicle. The high utilization (12+ hours per day) and relatively low throughput results in a much higher multiplier of 5.4 hphr/ton.

Landfill Calibration

Landfills, due to the number of facilities surveyed and in order to determine the existence of economies of scale, are treated slightly differently. Similar to the earlier treatments, reported hours per vehicle are aggregated for each site into a single number representing the disbursement of machine activity throughout the year. With enough points gathered to span the range of operations, regression is utilized to fit the data across the range of Q_{LF} to predict M_{LF} . In this way, a stationary, yet non-uniform estimate of machine activity can be produced for any candidate facility using the landfill's annual disposal tons as Q_{LF} . This means that, if a statistically significant regression is found, the machine activity rate depends on the size of the facility.

Orange County Waste and Recycling (OCWR) gave us fleet inventories and average annual hours for the three active landfills they manage: Frank R Bowerman (FRB), Olinda Alpha (OA), and Prima Deshecha (PD). In total, their fleet included 119 vehicles across 17 types. For

each model, the rated horsepower was found through manufacturer specifications and the average hours provided by OCWR. Load factors came from EMFAC.

Riverside County’s Department of Waste Resources also provided data for their four landfills. They gave us access to their full database of vehicle information and telematics for their fleet of 57 vehicles for 4 quarters in 2023-2024. These were separated into fleets at separate facilities and total hphr were calculated similarly as above.

Chicago Grade landfill in San Luis Obispo county is considered small at about 91,000 tons per year (near 40th percentile). They utilize just one vehicle per type: compactor, trash dozer, water truck, excavator, dump truck. They provided average weekly hours for each vehicle which was extrapolated to total horsepower hours for a year of operations.

Table 4.5 Annual horsepower hours and tons disposed for respondent landfills

<i>Site</i>	County	Power/annum (hphr)	Disposal/Annum (tons)	Total (hphr/ton)
Frank R Bowerman	Orange	6,116,420	2,243,109	2.72
Olinda Alpha	Orange	5,538,732	2,126,774	2.60
Prima Deschecha	Orange	3,876,444	626,281	6.19
Chicago Grade	San Luis Obispo	671,616	90,915	7.39
Lamb Canyon	Riverside	3,051,687	613,506	4.97
Badlands	Riverside	2,732,760	914,791	2.98
Blythe	Riverside	441,821	25,983	17
Oasis	Riverside	38,926	621	62.7

Results for the eight respondent landfills are displayed in Table 4.5. Frank R Bowerman and Olinda Alpa have the highest overall hphr generated due to their large disposal quantities in excess of two million tons per year. They also have lowest hp usage per ton due to apparent economies of scale. Interestingly, it seems that FRB’s extensive usage of scrapers rather than excavator/dump-truck for dirt-moving increases its energy disbursement over OA’s even though

OA disposes of slightly more waste. The smaller landfills exhibit much higher power needs per ton indicating that their operations benefit less from economies of scale.

With the addition of several facilities with less than one hundred thousand tons of MSW disposal annually, the machine-power needs over the full range of operation size was revealed to be highly nonlinear.

The relationship between landfill-specific equipment activity and annual disposal is modeled using a power function of the form $r_{hphr}(d) = \beta_0 d^{\beta_a}$. This functional form is consistent with operations literature describing specific requirements as a function of process throughput (Savage et al., 1986). The model is estimated by applying linear regression to the base-10 logarithms of r_{hphr} hphr/ton and disposal d tons for all 8 observations.

$$\log(r_{hphr}(d)) = -0.385 + 2.88d \quad (4.6)$$

This results in much better model fit across the range of landfill sizes which spans multiple orders of magnitude. The estimation yields a very good statistical result, with an R-squared of 0.96 and highly significant p-values for F- and t-statistics.

The resulting model provides an excellent fit across the observed range of landfill sizes, which spans multiple orders of magnitude. Estimation yields an (R^2) value of 0.99, with both the overall regression (F-test) and parameter estimates (t-tests) exhibiting high statistical significance as displayed in Table 4.6.

Table 4.6 Landfill Heavy Vehicle Specific Energy Model Significance

Model	Observations	F Stat.	R-squared	Coefficient	Coefficient
2	8	187.6****	0.99	Intercept	2.88****
				X Variable 1	-0.385****

The power-law formulation is also theoretically consistent with the expectation of economies of scale in landfill operations, whereby larger facilities require less equipment activity per ton of waste processed. Moreover, the model is straightforward to implement computationally while producing stable and highly significant parameter estimates. Accordingly, this log-log specification is adopted for subsequent analysis.

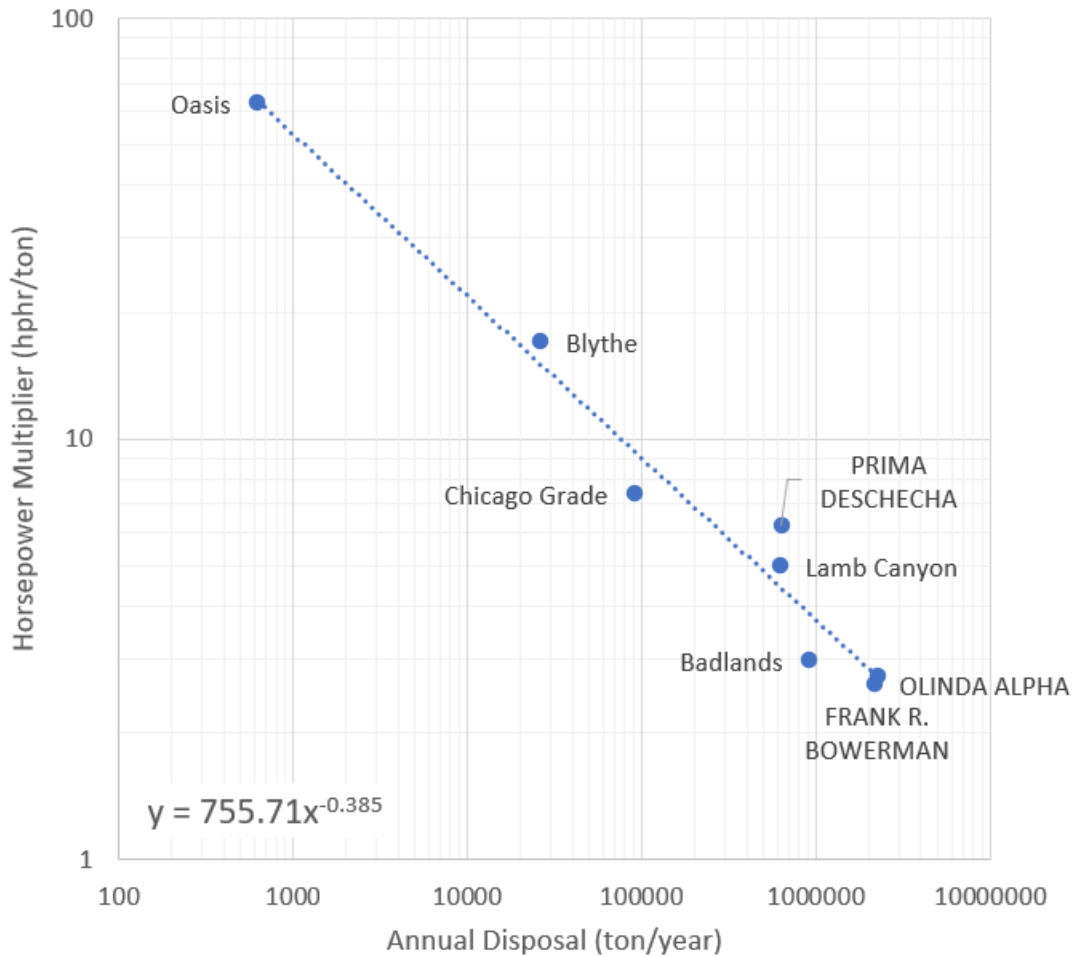


Figure 4.4 Nonlinearity (Economies of Scale) in Landfill Heavy Equipment Usage

$$M_{LF} = 755.71 * Q_{LF}^{-0.385} \quad (4.7)$$

The math behind the power model (equation 4.7) offers a constant elasticity interpretation, implying that a 1% increase in tons handled corresponds to a 0.385% decrease in machine hours *per ton* across all landfill sizes. This confirms strong economies of scale in landfill heavy equipment operations.

Fleet Simulation

Because detailed fleet inventories are generally unavailable for waste management facilities, estimation of heavy off-road vehicle emissions requires reconstruction of a plausible equipment population capable of satisfying observed facility activity. Unlike on-road transportation, where activity can be represented through vehicle miles traveled (VMT), off-road emissions modeling in the California waste sector must instead rely on equipment operation characterized through horsepower-hours (hphr). Within the California Air Resources Board (CARB) EMFAC off-road modeling framework, emissions are fundamentally determined at the vehicle level as a function of equipment horsepower, age, accumulated machine hours, annual activity, and load factor. However, statewide reporting systems such as the Recycling and Disposal Reporting System (RDRS) provide information regarding annual throughput and material handling, rather than the number, size, or age of machines operating at individual facilities. Consequently, a stochastic fleet synthesis procedure was developed to bridge facility throughput observations with the vehicle-level inputs required by EMFAC.

The necessity of simulating a fleet population arises from the fact that multiple combinations of equipment can plausibly satisfy the same operational demand. For example, a facility handling a given annual tonnage may operate a small number of high-horsepower wheel loaders, a larger number of lower-horsepower machines, or some combination thereof. While these alternative equipment populations may satisfy similar throughput requirements, they can

produce materially different emissions outcomes, particularly for criteria pollutants such as nitrogen oxides (NO_x) and particulate matter (PM), whose emission rates depend strongly on model year and deterioration. As a result, assignment of a deterministic “representative vehicle” would obscure uncertainty and fail to represent the heterogeneity of real-world facility operations. Instead, facility-level emissions are estimated through repeated stochastic simulation of feasible vehicle populations.

The simulation framework begins by estimating facility power demand using reported annual throughput. A deterministic horsepower-hour intensity parameter, (M_f hphr/ton), calibrated from transfer station operator interviews, is multiplied by annual throughput to estimate the total horsepower-hours required to sustain operations. For transfer station wheel loaders, an initial estimate of ($M_f = 1.8$) hphr/ton was adopted based on observed equipment utilization at the Irvine Transfer Station. Required horsepower-hours are then translated into a nominal equipment demand using assumptions regarding annual operating hours and load factor. Annual equipment utilization is initialized using a deterministic value of 2,250 hours per machine-year, representing approximately five to six days of operation per week over a full year. A wheel loader load factor of 0.37, obtained from EMFAC off-road defaults, is used to convert rated horsepower into average time-weighted horsepower expenditure during operation.

Given the estimated operational demand, a feasible vehicle fleet is synthesized through stochastic population modeling. Equipment is categorized into small, medium, and large horsepower classes defined by bounded horsepower ranges. A factorial design is used to generate candidate fleet configurations consisting of varying combinations of these equipment classes. The resulting candidate populations represent feasible technology bundles capable of supporting

facility operations. For each configuration, a nominal fleet power is calculated using representative horsepower values associated with each equipment class.

A geometric nearest-feasible-solution heuristic is then used to select candidate fleet populations capable of reasonably satisfying facility demand. Candidate fleet configurations are represented as points in multidimensional space, while the facility's estimated power requirement defines a constraint surface. Distances between candidate fleet configurations and the target demand are evaluated, and satisfactory fleet configurations are identified based on proximity thresholds. Where multiple feasible configurations exist, one configuration is randomly selected to inject some stochastic variation between otherwise similar facilities (or the same facility across runs). This procedure allows the model to maintain operational realism while avoiding over-specification of fleet characteristics unsupported by available data.

Following selection of a fleet configuration, detailed vehicle populations are simulated by stochastically assigning attributes to each machine. Horsepower is sampled from a uniform distribution bounded by the selected equipment class range, while vehicle age is sampled from an empirical probability mass function representing the age distribution of off-road equipment (assumed to match the broader construction and industrial category). Vehicle age is particularly important because EMFAC emissions rates are sensitive to historical changes in emissions standards and equipment deterioration over accumulated operating hours. Older machines are subject to less stringent emissions controls and exhibit higher deterioration-adjusted emissions rates, particularly for NO_x and PM emissions. Lifetime machine hours are estimated as the product of annual operating hours and equipment age, enabling deterioration effects to be incorporated into emissions estimation.

For each simulated vehicle, horsepower-hours are calculated as the product of rated horsepower, annual operating hours, and load factor. EMFAC off-road emission factors are then applied as a function of vehicle age, horsepower bin, forecast year, and pollutant-specific deterioration and fuel correction terms. Emissions are estimated at the individual vehicle level and aggregated to facility, county, and statewide totals. Because the vehicle population itself is simulated rather than directly observed, repeated Monte Carlo realizations are used to characterize uncertainty and produce mean estimates, variances, and confidence intervals for emissions outcomes.

Importantly, the framework is modular and extensible. Parameters governing equipment utilization, horsepower-hour intensity, age distributions, fleet selection rules, and even emissions rate models may be substituted as improved empirical information becomes available. In this way, the stochastic fleet synthesis procedure functions not merely as a source of uncertainty, but as a necessary inferential mechanism for reconstructing latent equipment populations required to operationalize facility-level off-road emissions modeling within a statewide waste management context.

4.3.3.3 Sorting Out Success Part 1: MRF Operations and Energy Usage

In addition to heavy vehicles, material recovery facilities (MRFs) have substantial emissions related to electricity usage. These facilities have numerous electricity-demanding mechanical unit operations including rotating screens and electromagnets as well as computers, controllers, compressors, and conveyance which work together to concentrate desirable recyclable commodities and reject contamination into distinct outflow streams.

Despite the increased facility complexity, the same general emissions quantification framework of activity estimates times emission rates is applied. To characterize activity, a separation process model is used to estimate recovery rates and throughput of unit processes. Then, electricity consumption and associated emissions are calculated.

MRF Unit Operations

The variety of MRF operations make declaring the presence, configuration, and ordering of the units uncertain, much less how much waste is flowing through each unit, which is subject to the composition of inflow material and sorting efficiency of each component. This uncertainty in composition, recovery, and purity outcomes coupled to process configuration is one gap that my conditional simulations address on the way to estimating electricity demand.

Modern material recovery facilities separate commingled recyclables into marketable commodity streams through a sequence of mechanical, magnetic, and sensor-actuator technologies. The unit operations exploit various physical properties to separate targeted waste resources for recycling. Although individual facilities vary considerably in equipment selection and process layout, most modern MRFs employ a broadly similar set of processes. Disc screens separate cardboard and paper. A so-called glass breaking screens separates metal and plastic containers from glass. Add a magnetic belt separator to capture iron and steel. Add eddy current after that to capture aluminum beverage cans. And finally, mixed plastic streams are separated by relatively high-tech sensors: these include optical (cameras trained on the inflowing belt) and near-infrared reflectance (which detect chemical structures, not color). Computers process the image and coordinate actuation by air knife to sequester ejection targets. These may be positive or negative sorting depending on the units' roles in the overall design (process parameters c,r,p).

In practice, MRF design represents a balance between material purity, commodity recovery, throughput, and operational reliability. Sorting systems are not typically optimized to maximize purity or recovery independently. Rather, facilities operate within a satisficing framework in which commodity specifications, market requirements, throughput targets, and avoidance of downtime collectively determine operating conditions. Managers indicated that maintaining stable operation and minimizing equipment downtime often outweigh marginal improvements in separation performance.

For the purposes of this research, the complexity of modern MRF operations is represented using a simplified process-flow model consisting of four separation stages, blocked as series of binary separations, as indicated in Figure 4.5. Disc screening recovers fiber. A glass breaker diverts glass and residual fines. Magnet and Eddy separators pull metal from the stream. And plastic is recovered by optical-airknife. The remainder (residual) flows to the refuse stream.

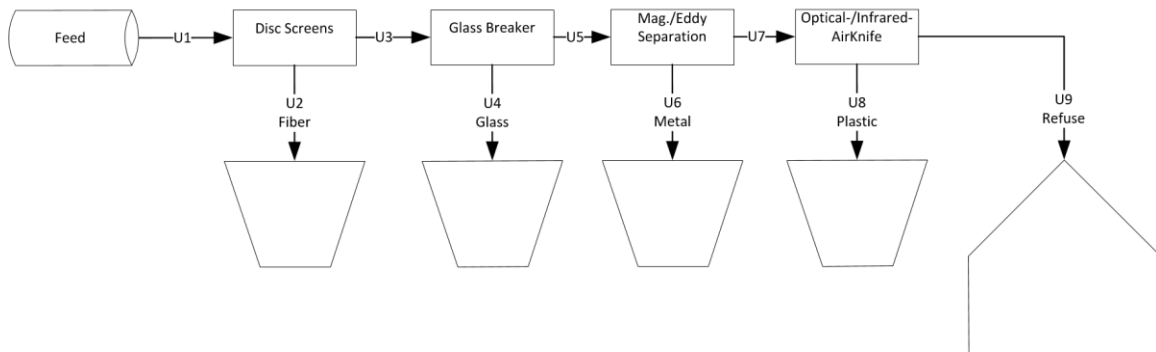


Figure 4.5 Block Diagram of MRF Process Flow

This mathematical abstraction captures the dominant material-specific separation mechanisms observed across modern MRFs while avoiding facility-specific details that are

unavailable at statewide scale. The resulting framework provides a practical representation of material recovery operations suitable for emissions estimation and system-level modeling.

Furthermore, it reduces the number of parameters.

Separation Process Flow Model

Traditional engineering process modeling provides a mathematical framework for defining the process as a system of equations. With sufficient specification per degrees of freedom, determination of all system parameters can be achieved. I will present and apply a fairly simple collection of balancing equations centered around the system illustrated in Figure 4.5.

For a single inflow of mixed recyclables, the following transfer equations describe the system. X indicates the total inflow of a mixed waste stream which consists of $i+1$ parts where i is the number of commodities targeted for separation. The ‘+1’ is for the contamination fraction denoted “Refuse”. The fraction of commodity i in X is denoted c_i and the contamination rate is c_{Refuse} . These are constrained to sum to 1 (per equation 4.11).

$$\begin{array}{c}
 X \\
 \text{Basis}
 \end{array}
 \begin{array}{ccccc}
 & F & G & M & P & Ref
 \end{array}
 \begin{array}{c}
 \mathbf{R}_{io}^T \\
 \\
 \\
 \\
 \\
 \end{array}
 \begin{array}{c}
 c_i \\
 \\
 \\
 \\
 \\
 \end{array}
 =
 \begin{array}{c}
 \mathbf{Y}_o \\
 \\
 \\
 \\
 \\
 \end{array}
 \quad (4.8)$$

$$\begin{array}{c}
 \text{fiber} \\
 \text{glass} \\
 \text{metal} \\
 \text{plastic} \\
 \text{refuse}
 \end{array}
 \begin{bmatrix}
 r_{fF} & 0 & 0 & 0 & r_{fR} \\
 0 & r_{gG} & 0 & 0 & r_{gR} \\
 0 & 0 & r_{mM} & 0 & r_{mR} \\
 0 & 0 & 0 & r_{pP} & r_{pR} \\
 r_{rF} & r_{rG} & r_{rM} & r_{rP} & r_{rR}
 \end{bmatrix}^T
 \begin{bmatrix}
 c_f \\
 c_g \\
 c_m \\
 c_p \\
 c_{ref}
 \end{bmatrix}
 =
 \begin{bmatrix}
 Y_F \\
 Y_G \\
 Y_M \\
 Y_P \\
 Y_{Ref}
 \end{bmatrix}$$

$$\mathbf{0} \leq \mathbf{R}, \mathbf{c}, \mathbf{p} \leq \mathbf{1} \quad (4.9)$$

$$\mathbf{R}\mathbf{1} = \mathbf{1} \quad (4.10) \text{ row sum of } \mathbf{r} \text{ constraint}$$

$$\mathbf{1}^T \mathbf{c} = \mathbf{1} \quad (4.11) \text{ col sum of } \mathbf{c} \text{ constraint}$$

$$\mathbf{1}^T \mathbf{p} = \mathbf{1}^T \quad (4.12) \text{ col sum of } \mathbf{p} \text{ constraint}$$

$$\mathbf{1}^T \mathbf{Y} = X \quad (4.13) \text{ Mass balance constraint}$$

$$X(\mathbf{c}\mathbf{1}^T) \odot \mathbf{R} = \mathbf{P} \odot (\mathbf{1}\mathbf{Y}^T) \quad (4.14) \text{ Defining element wise product and division as } \odot \text{ and } \oslash, \text{ respectively}$$

Y indicates the total outflow which is also divided into $i+1$ parts, except these parts are represented as separate streams, not as a single, mixed stream. Mass balance is enforced – that is, X is calculated from reported outflows per equation 4.13.

Recovery rate, r , is defined as the ratio of how much material was sorted into the correct outflow stream per the total inflow quantity of that material. If recovery rates are less than 1, some commodity material is lost in the refuse stream. This is distinct from purity, p_i , which measures that fraction of an outflow stream that is the target material. When purity is less than 1 for one or more commodities, it is assumed that only refuse makes up the contaminated fraction of a given outflow.

These simplifying assumptions allow certain parameter values to be constrained. For example, my formulation assumes only refuse contaminates product streams. This allows the assigning of 12 zeroes into the recovery (eq 4.8 and 4.14) and purity matrices (eq 4.14). Purity $\mathbf{P}$ is defined here as a special case of composition. On the one hand, it is more complicated because it stores multi-stream compositions, but on the other, is simplified to just two components for commodity streams (desired and undesired product). Constraints on row and column sums of recovery (eq 4.10) and composition (eq 4.11, 4.12) provide a number of equations relating parameters and ultimately specify one degree of freedom per material stream each.

But the real breakthrough comes with realizing that when purity is no longer binding ($p < 1$ for any material/stream), this means $r_{rR} < 1$ such that some refuse material contaminates the outflow. Similarly, if $r_{i,i} < 1$, then the corresponding material in the refuse outflow must satisfy $c_{iR} > 0$. In fact, for each commodity and refuse stream, observed outflows and estimated purities define a minimum inflow composition required to achieve a feasible recovery rate.

Consequently, a tradeoff emerges between assumed inflow composition and achievable recovery. Ceteris paribus, increasing the fraction of a commodity in the inflow decreases the implied recovery. Depending on the perturbation scheme, this may correspond to an increase in recovery slightly for one or more of the others. This relationship creates a feasibility constraint that defines the minimum and maximum amount of each commodity in the inflow stream.

Previous Monte Carlo implementations of this framework implicitly identified these feasibility bounds by repeatedly sampling candidate compositions and retaining only feasible solutions. Examination of the resulting distributions revealed that feasible solutions terminate at distinct lower bounds corresponding to the minimum composition required to produce the observed purified outputs. These bounds can therefore be derived analytically rather than solely through simulation.

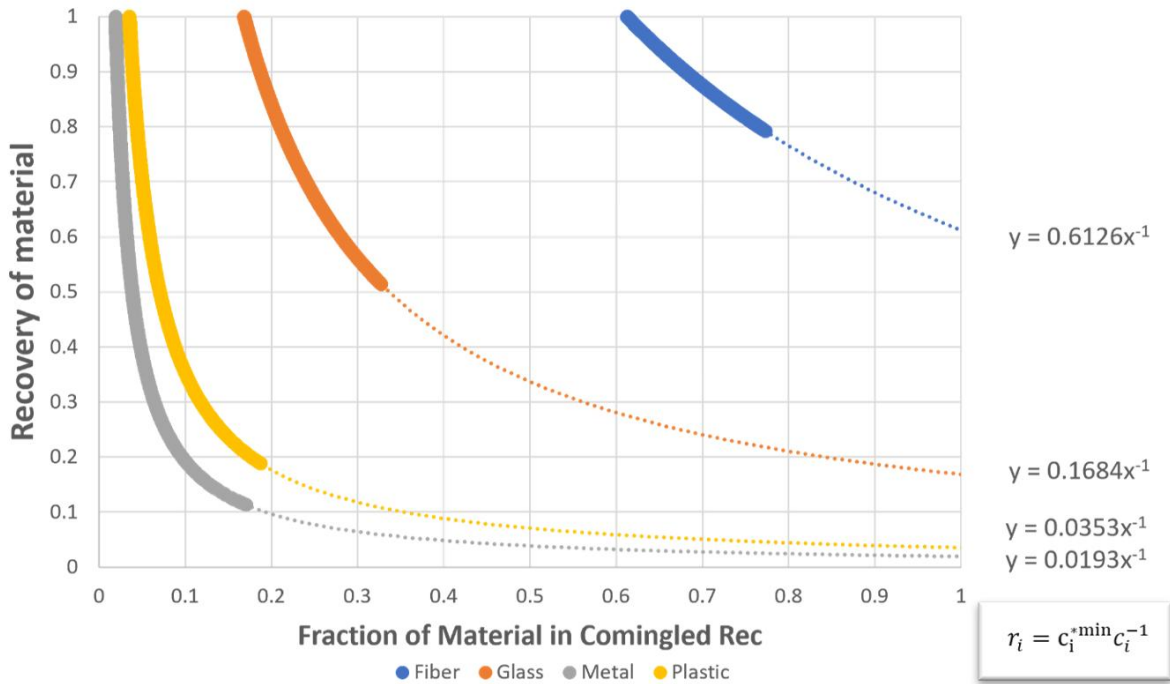


Figure 4.6 Feasible Recovery and Inflow Composition Sub-Space

Solving eq 4.14 for $\mathbf{r}$ and then substituting eq 4.9 produces the minimum compositions:

$$\mathbf{c}\mathbf{1}^T \geq \mathbf{c}^{*min} = (\mathbf{P} \odot (\mathbf{1Y}^T)) \frac{1}{X} = \begin{matrix} & \mathbf{P}_{io} & & \mathbf{1Y}_o^T/X \\ \begin{bmatrix} c_{fF}^{(p)} & 0 & 0 & 0 & c_{fR} \\ 0 & c_{gG}^{(p)} & 0 & 0 & c_{gR} \\ 0 & 0 & c_{mM}^{(p)} & 0 & c_{mR} \\ 0 & 0 & 0 & c_{pP}^{(p)} & c_{pR} \\ c_{rF}^{(1-p)} & c_{rG}^{(1-p)} & c_{rM}^{(1-p)} & c_{rP}^{(1-p)} & c_{rR} \end{bmatrix} & \frac{1}{X} \begin{bmatrix} Y_F \\ Y_G \\ Y_M \\ Y_P \\ Y_{Ref} \end{bmatrix}^T \end{matrix} \quad (4.16)$$

The outcome of this system can be reduced to the most restrictive constraint. This will obviously not include the 0 elements of $\mathbf{P}$ and in a simple binary separations problem, strictest constraints generally lie on the diagonal. Any minimum composition above 0 constrains the others' components with an accumulated ceiling due to the compositional unit sum. The difference in these values is the same within each material balance. Thus, every material has a minimum and maximum value constraining their contribution toward the inflow composition. These critical values are defined as vertices of a feasible space. For example, the constrained system could increase the composition of one component by c_s , if the rest stayed at the minimums.

In order to explore the feasible space, all materials are assumed to be 1:1 substitutes on a mass basis and a simplex-constrained factorial mixture matrix was developed. This allocates slack in different proportions to produce numerous $(\sum_{k=1}^5 \binom{5}{k}) = 2^5 - 1 = 31$ linear combinations of feasible initial points. These were crossed with every item in the sample to generate feasible allocations of compositional slack per facility. Many of these still result in full recovery – 0 slack on this constraint is related to systematic data issues to be discussed.

Additional Data

Having established the mathematical framework, the remaining task is to close the system of equations by introducing sufficient empirical constraints to resolve the remaining degrees of freedom.

As suggested many times above Commodity yields (Y) and product purities (p) are specified first because they can be estimated directly from RDRS material-flow records and CalRecycle MRF characterization studies. By treating these quantities as exogenous inputs, the dimensionality of the solution space is substantially reduced, allowing the remaining endogenous variables to be identified through the model constraints.

RDRS data for the MRF sample is aggregated into a coarser, set of materials than that applied in Chapter 3 to coincide with the simplified process diagram. Fiber, Glass, Metal, Plastic, and Refuse are denoted by the subscripts f , g , m , p , and r , respectively.

Next, I incorporate commodity purity estimates derived from another privileged dataset from CalRecycle, their MRF survey. CalRecycle conducted outreach, phone surveys, and in some cases, in-person surveys and field work which included waste characterization of multiple sample points in the MRF processes. The survey includes multiple MRFs throughout the state, multiple sampled streams, and repeated observations (460 samples across 23 facilities), providing a robust empirical basis for estimating commodity purities. Because composition data are constrained and compositional in nature, an isometric log-ratio (ILR) modeling framework (see Appendix D.1) is employed to characterize the joint distribution of material fractions while respecting the dependency structure among components.

I remapped streams and materials to reflect label-matching criteria employed above and calculated purity for each of the streams indicated in Figure 4.7. The mean, upper- and lower-95% CI estimates of purity are indicated for each stream. Most of the commodities (fiber, plastic, and metal categories) have purity estimates above 95%. However, cardboard and paper have more apparent variation. This is consistent with our conversations with operators who target fiber purity in the low 90s but may pursue higher-grade contracts periodically. Glass is notably lower (mean of 80% with a fairly wide range) which is indicative of the glass-breaker screen's poor selectivity. This is accepted because glass is typically transported to a glass processor who sorts by color to produce uniform cullet for remanufacturing. The MSW (aka Refuse or residual) stream has the lowest mean “purity” at 60% and a lower bound at about 45%. This means that a significant fraction of single stream material is either non-recyclable materials or are not recovered.

Table 4.7 Operational Purity Values

Material in Material Stream	$P_o^{(p)}$
Fiber	0.96
Glass	0.80
Metal	0.95
Plastic	0.98
Refuse	0.62

In any case, the purity estimates for the commodities are very high. Their averages within groups are nearly identical, which supports “blocking” them. The averages in Table 4.7 will be utilized in the process flow model calculations.

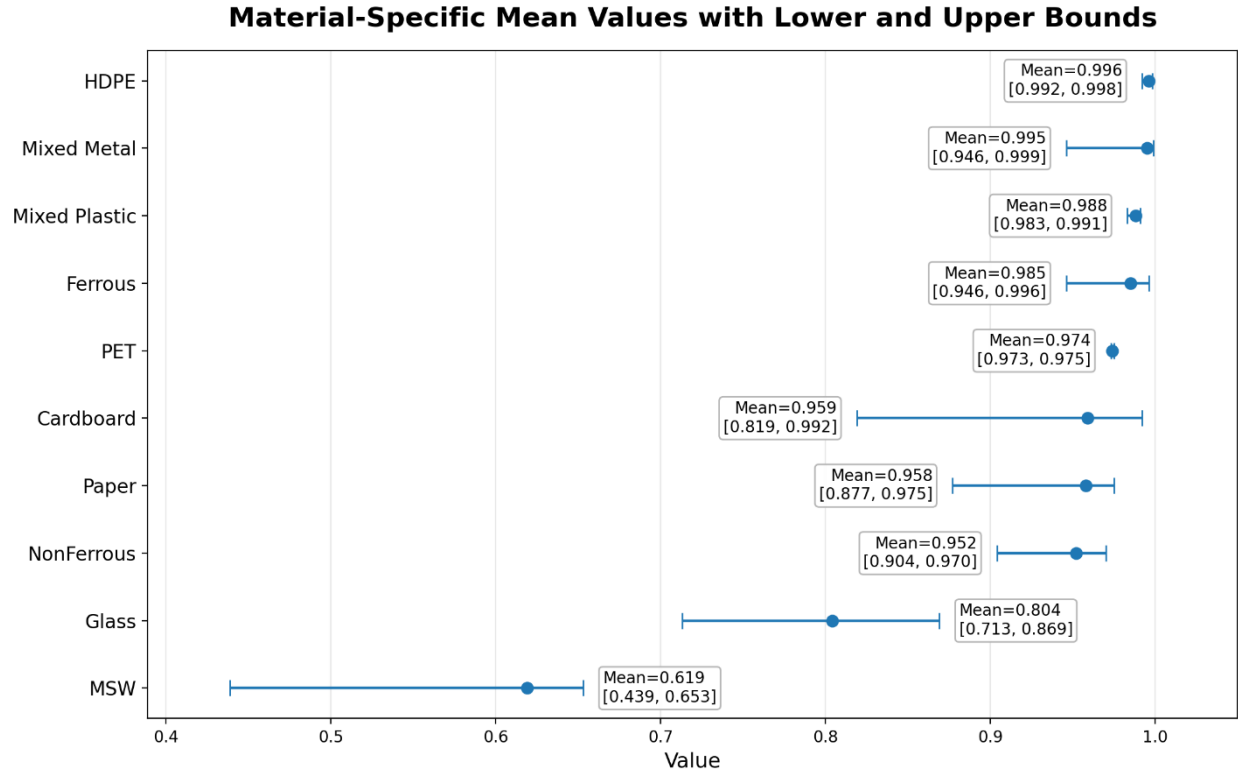


Figure 4.7 Material Stream Purity Estimates

Estimation and the Objective Function

In seeking an estimator or statistic or to either constrain or optimize, I utilize the fact that the total amount of rejected outflow stream aka residuals, Y_R , must equal the entering quantity of refuse sorted correctly ($Xc_r r_{rR}$) plus commodities which are sorted incorrectly ($Xc_i(1 - r_i)$).

Substituting equation 4.13 and rearranging:

$$p_{Refuse} = 1 - X \sum_i c_i (1 - r_i) \quad (4.17)$$

While I personally find the idea of specifying the “purity” of the refuse stream slightly strange, the data from CalRecycle does support its estimation, but the implications of its complement (lost commodity i.e. poor recovery) should not be claimed with certainty. As an initial condition

this is fine but I hypothesized that the problem behaves better when this constraint is allowed slack. So, rather than introduce it as a binding constraint, I introduced it to the objective function as another SSE term between $c_i(1 - r_i)$ and $\frac{Y_{Refuse}(1-p_{ref})}{X}$. This seems to provide better convergence, but does not guarantee replication of the p_{ref} input. A feedback approach could likely solve this but was not implemented as the problem begets feasible solutions, the primary goal, and the imbalance is only detectable when exhaustively verifying the full system of equations.

The more dogmatic use of equation 4.17 is as a binding constraint. This seems to make convergence more difficult but does provide stability where the reported data, and its perturbation by the factorial design, are close to the distribution measured by CalRecycle technicians. Then again, in those cases, the prior formulation works equally well.

The primary optimization target is that the model output quantity $c_i r_i$ matches the target $c_i^* r_i^*$ prescribed by the factorial mixture. The full mathematical program is described by eq 4.32.

$$\text{Min } SSE(c_i^* r_i^*, c_i r_i) \quad [\text{eq 4.18}]$$

Decision Variables: c_i, r_i Subject to: equations 4.7 through 4.17

Process Specific Energy

The optimization problem is run for all MRF facilities, for all mixture perturbations. Some instability is noted in certain facilities (those with free parameters aka $Y_o = 0$) but the results are pooled and an average unit-process-level flow is generated. This fraction is applied to machine power ratings to calculate process level specific energy as indicated in Table 4.8.

The baseline emissions intensity (kg CO₂/kWh) is from the CAPCOA '21 handbook which provides an average for the state as well as by energy provider (Caltrans, 2021).

Table 4.8 MRF Separation Process Specific Energy and Emissions

Power Rating (kW)	Load Factor	Unit Throughput per Feed Basis $\left(\frac{ton_{Unit}}{ton_{Process}} 100\%\right)$	Separation Average R_{eff}	Unit Specific Energy $\left(\frac{kWh}{ton_{Unit}}\right)$	Process Specific Energy $\left(\frac{kWh}{ton_{Feed}}\right)$	Emissions Intensity $\left(\frac{kgCO_2e}{ton_{Feed}}\right)$
Drum Feeder			Feed			
15.00	0.75	100.00%	0.00%	12.00	12.00	3.61
Disc Screens (Block)			Fiber			
13.65	1.00	100.00%	35.00%	13.65	13.65	4.10
Glass Breaker (Block)			Glass			
158.20	1.00	65.00%	12.00%	158.20	102.83	30.91
Electro-Mechanical (Block)			Metal			
10.00	1.00	53.00%	6.00%	10.00	5.30	1.59
Optical (Block)			Plastic			
20.80	1.00	47.00%	5.00%	20.80	9.78	2.94
Aggregate Process-Specific Energy and Emissions per ton of feed.					143.6	43.2

4.3.3.4 Baling

Baling is a mechanical densification process used in recycling and waste management to compress loose materials into dense, standardized bundles for efficient storage, handling, and transport. In practice, baling typically occurs after sorting and contaminant removal (i.e. at the end of a material recovery process). The recovered materials are fed into a baler, which applies high hydraulic pressure to compact the material. The compressed block is then secured using wire, steel strapping, or plastic ties to maintain its shape during handling and shipment. In the context of this research, I assume that certain material outflows (aluminum, cardboard, paper, and plastics) are all baled as they exit a MRF.

Baling machines are powered by electricity and follow a power profile similar to the “Assumed Continuous Power Demand Profile” presented in Figure 4.8. In the compression phase, the profile ramps up to the full power load factor as the hydraulic ram presses material into a smaller volume. Once the bale is secured by several loops of baling wire, the ram is returned to the starting position and the bale ejected. The profile ramps up to the maximum power output by about halfway through the cycle, lingers briefly while wire is secured, and falls at a slightly faster rate as the machine returns to its starting position. Some balers have more than one ram

Power demand and therefore carbon emissions from the electricity are determined by this profile. Calculating the integral of the profile, normalized as shown, results in the load factor. Multiplying by the cycle time and maximum power rating provides an estimate of power demand in kWh/cycle. The annual total emissions can be determined by multiplying by the number of cycles and the GHG per kWh of the electrical grid.

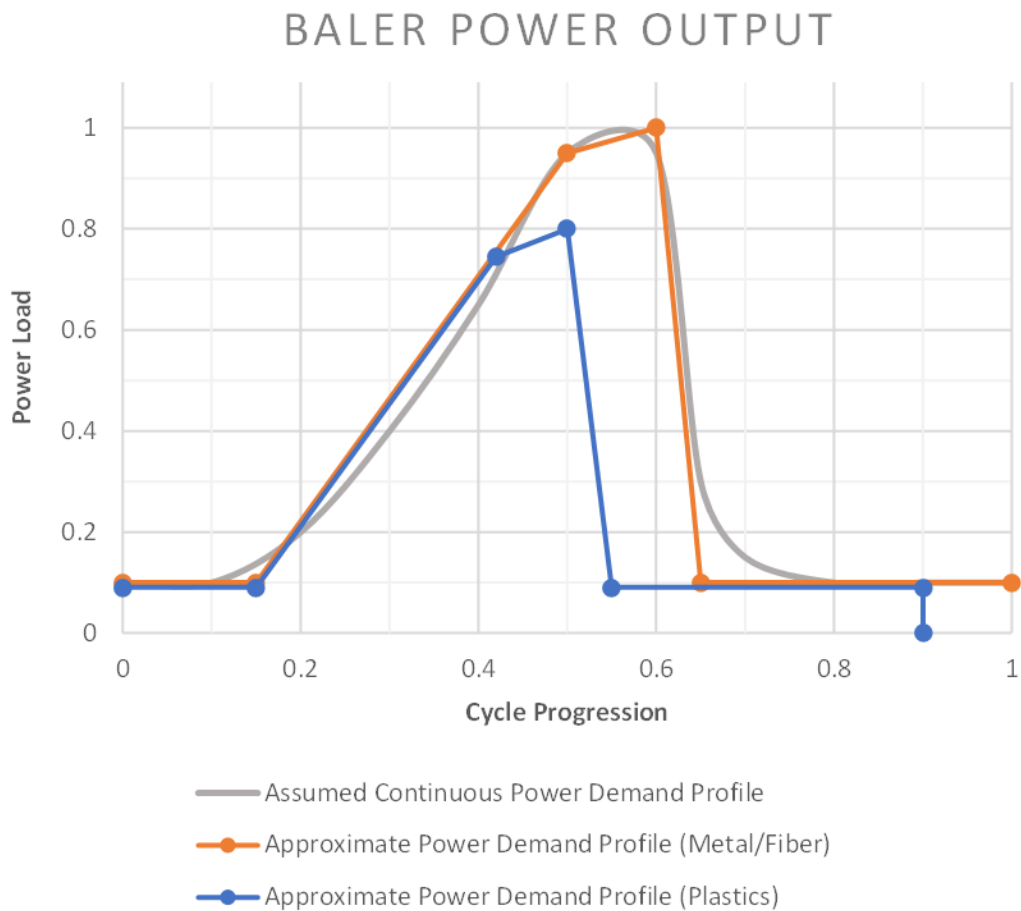


Figure 4.8 Baler Power Output Profiles per Cycle

My interview with the maintenance supervisor at WM Orange revealed some very interesting operational nuance related to differences in procedure by material. He noted that, when baling plastics, maximum compressive power is reduced by about 20% because the material elasticity is great enough to break the baling wire upon cycle completion if it is compressed further. Also, due to differences in density, different materials have different standard bale sizes and mass. Metals and fiber are usually finished at 1000 lb per bale, whereas plastics are 800 lb per bale even though the “cubes” are larger.

To calculate separate profiles for plastic vs other materials, the assumed power demand profile is simplified to be piecewise linear. Then, a profile for plastics baling is developed which is truncated in terms of power load (max of 0.8 instead of 1.0) and cycle time (10% decrease in cycle time because the max is reached sooner). Integration of these profiles is performed geometrically, resulting in load factors of 0.25 for plastics and 0.36 for other materials. I base other operational parameters on the conditions at WM Orange, including a nominal 60hp baler, 40 second average cycle time (1/90 hr/cycle), and the aforementioned bale weights. The 2021 statewide average GHG emission factor for electricity is taken from table E-4.3 of the 2021 CAPCOA handbook (CalTrans, 2021). Note that actually running 90 cycles per hour on one machine is not realistic, but the time spent filling and emptying of the machine need not be included in this annual calculation. Also, the reduction in cycle time for plastics is accounted for by the load factor integration. The resultant per-ton emission factors are presented in Table 4.9.

Table 4.9 Baling Emissions

Material	$Load\ Factor \times \frac{hp \cdot hr}{cycle} \times \frac{kW}{hp} \times \frac{g\ CO_2e}{kW \cdot hr} \div \frac{ton}{cycle}$					Result
Plastics	0.25	$\frac{60}{90}$	0.7457	195	0.4	$9.7 \frac{g\ CO_2e}{ton}$
Metal, Fiber	0.36	$\frac{60}{90}$	0.7457	195	0.5	$17.4 \frac{g\ CO_2e}{ton}$

These factors are multiplied by the reported outflows, in short tons, of recyclable commodities from MRFs.

4.4 Synthesis

Although the collection, transportation, and facility components are estimated using distinct methodologies, they are ultimately derived from a common emissions framework and are therefore integrated into a unified statewide inventory. Each component utilizes the modeling approach most appropriate for the available activity data: municipal collection emissions are estimated directly using EMFAC on-road inventory, post-collection transportation emissions are obtained through routing-based vehicle activity estimates coupled with EMFAC, and onsite facility emissions are generated through conditional simulation of equipment populations and activity. Despite these methodological differences, all results are standardized to a common reference year (2021), geographic framework (California Air Basin District subareas), and pollutant accounting framework, including both criteria pollutants and carbon dioxide-equivalent greenhouse gases. This harmonization permits direct comparison, aggregation, and attribution of emissions across all modeled components while preserving the engineering assumptions unique to each subsystem.

The remainder of this chapter applies the framework to the main components of California's waste and recycling infrastructure. First, statewide estimates for each method (collection, inter-facility transport, and on-site methods) are presented. Then, the system is broken up into subsystems, defined in terms of material and the facilities involved. This treatment sets up the analysis of offsets for waste stream handling by different pathways.

4.4.1 Statewide Emissions

System totals are presented in Table 4.10. The relative impact with respect to GHG statewide is heavily in favor of collection at 960,000 metric tons of GHG released. This is mainly because EMFAC assigns collection trucks very poor duty cycle characteristics. The results suggest that

collection trucks operate at an inferior fuel rate of 3 miles/gallon, more than twice the implied inter-facility transport fuel economy.

Table 4.10 Statewide Carbon Footprint by Source in Base Year ('21)

Measure\System	Collection	Inter-Facility Transport	Off-Road Vehicles	On-site Electricity (Separation)	On-site Electricity (Baling)
VMT (MVMT)	328	101	n/a	n/a	n/a
Fuel (gal)	96,438,700	13,362,522	3,782,257	n/a	n/a
Electricity (MWh)	n/a	n/a	n/a	386	0.02
GHG (MgCO ₂ e)	960,217	140,334	38,617	113	0.5

Figure 4.9 indicates the collection and inter-facility transport by county. This map shows that the high ratio of collection activity is less pronounced in more rural areas, where the on-road numbers begin to balance out the collection. This makes sense because more small transfer stations would be used and the distances for collection are much smaller due to small towns.

Similar patterns are visible in Figure 4.10 comparing criteria pollutants by county. The patterns of mobile source emissions echo general population presence. However, there are slight differences between them. Collection seems to have less diffuse distribution of mileage whereas the inter-facility transport estimates indicate elevated mileage in central California, suggesting that inter-county external/through trips are traveling the I-5 freeway.

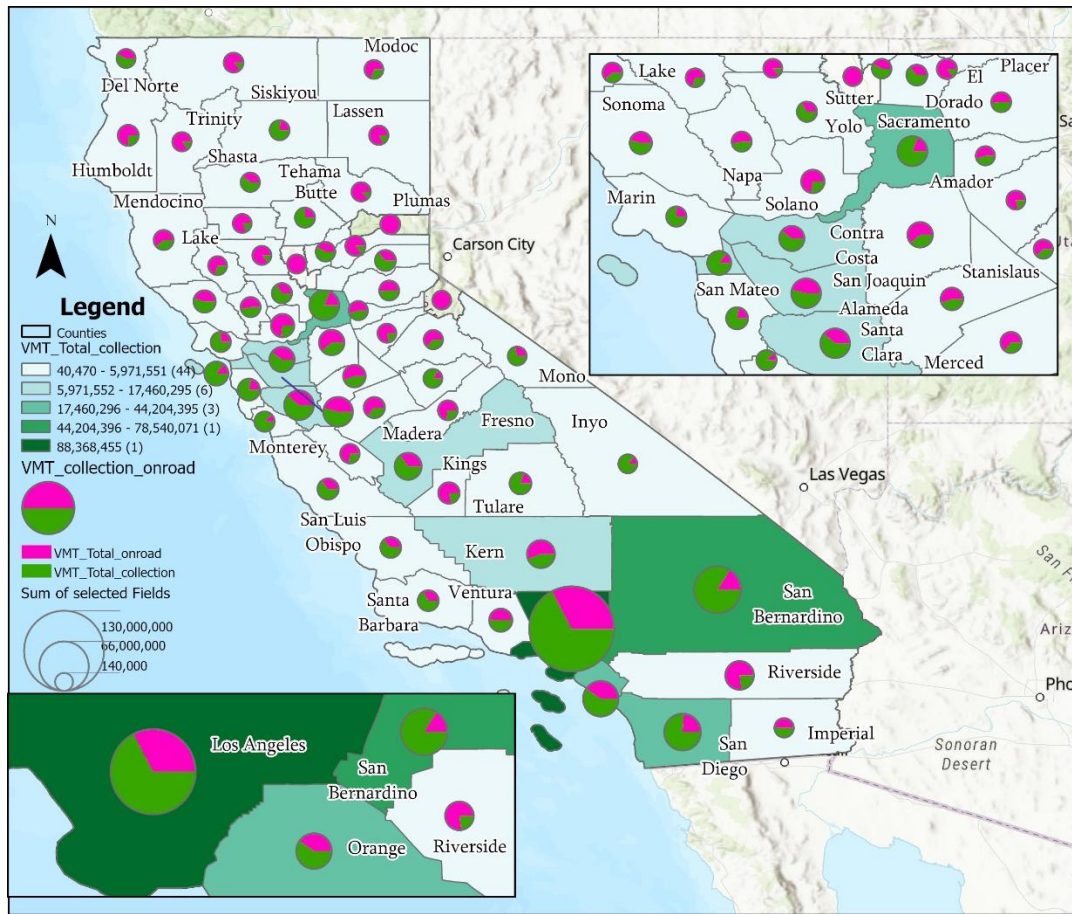


Figure 4.9 On-Road GHG Intensity Comparison

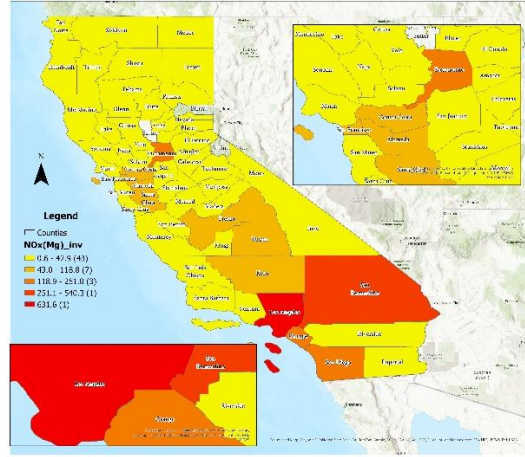
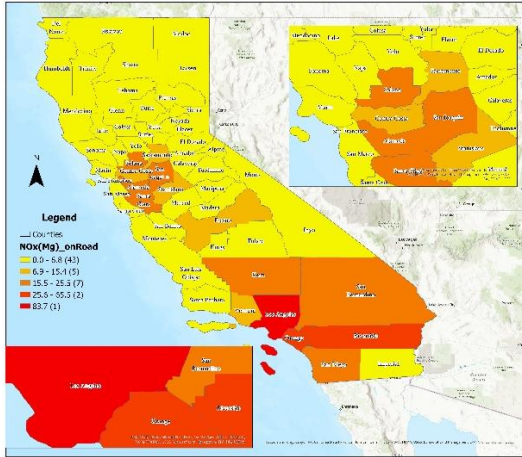
Off-road vehicles come in at a much lower overall emissions level. This is mainly because the model is only applied to a small number of facilities (MRFs). MRF electricity usage is significant, again considering the number of facilities involved in the calculation. Baling expenditure is essentially nonexistent due to the very fast cycle time and low load that I selected for the baling calculations. This suggests the parameters in section 4.3.3.4 (cycle time and load factor) need to be adjusted closer to Testa (2015).

Figure 4.11 shows the results of fleet simulation. These numbers have some variation due to the variates generated on fleet configuration (and rated horsepower),

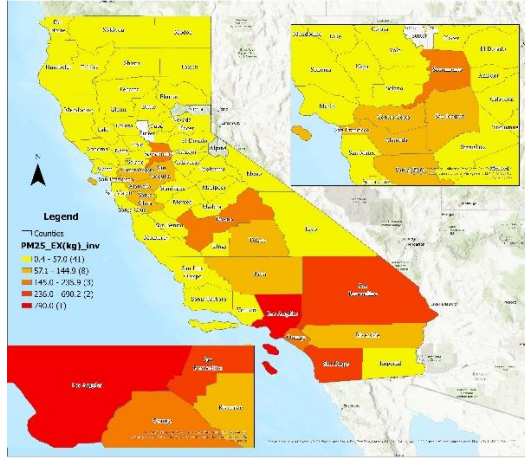
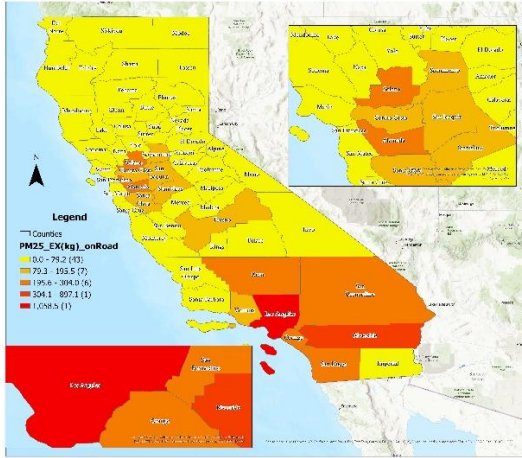
Post Collection Transport

SWCV Inventory

Nox



Combustion PM



Well-to-Wheels PM

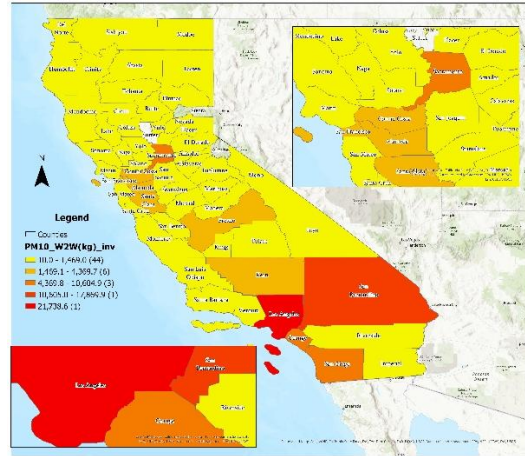
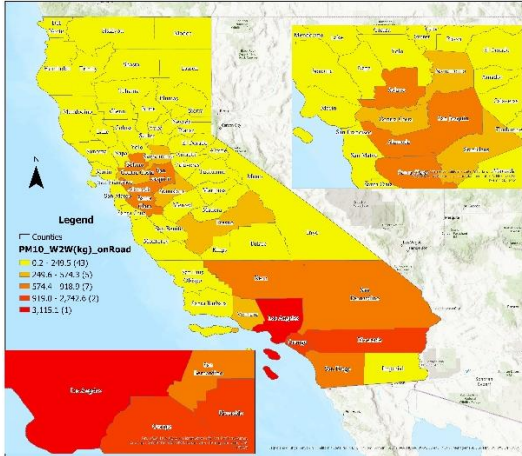


Figure 4.10 Distribution of NOx and PM for On-Road Activity

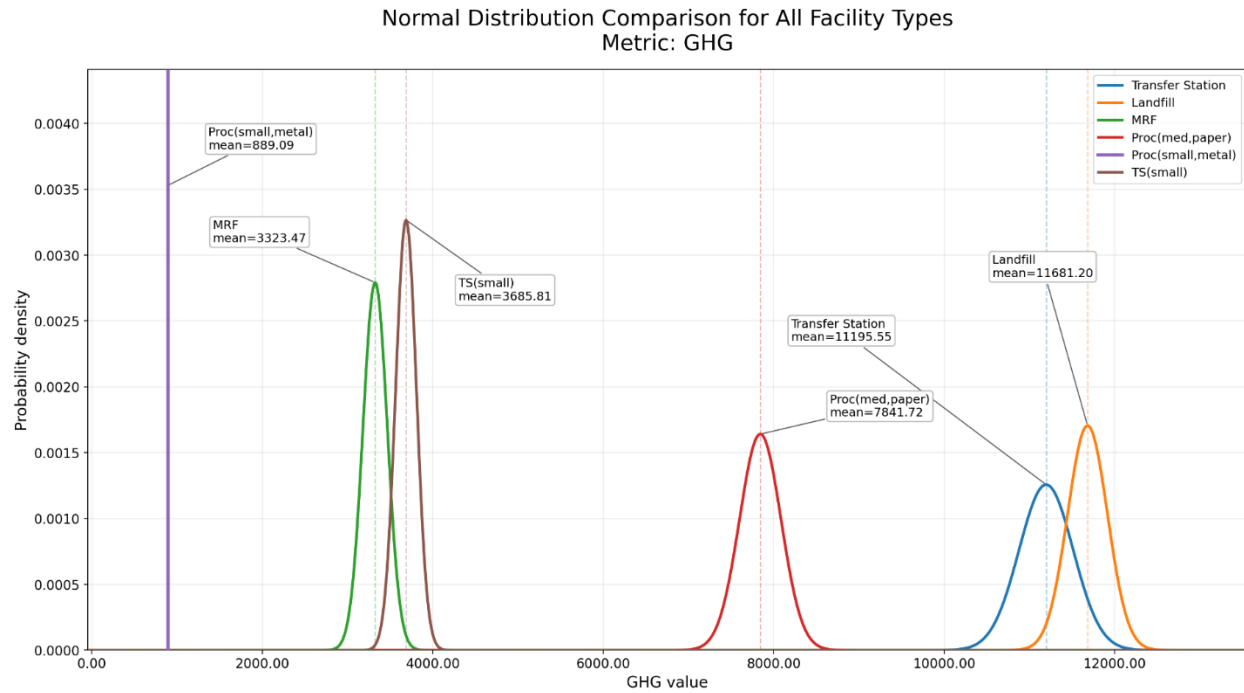


Figure 4.11 On-Site GHG Emissions Inventory

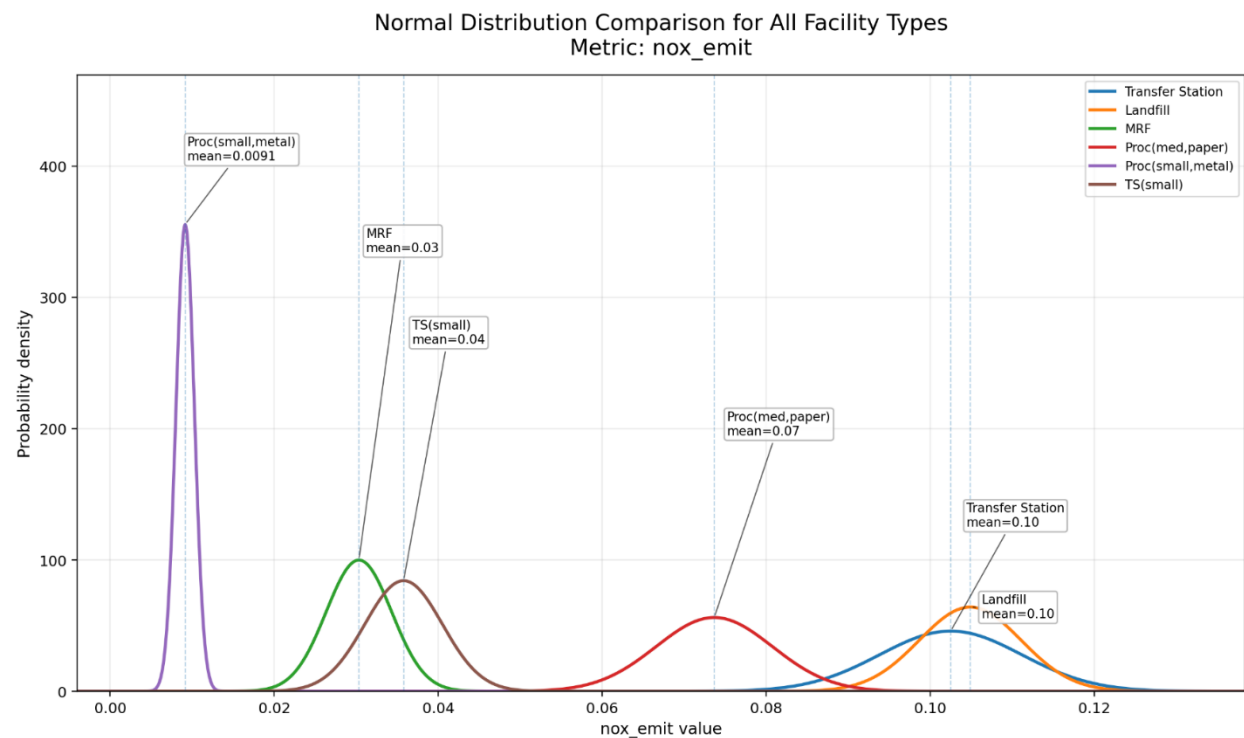


Figure 4.12 On-Site NOx Emissions Inventory

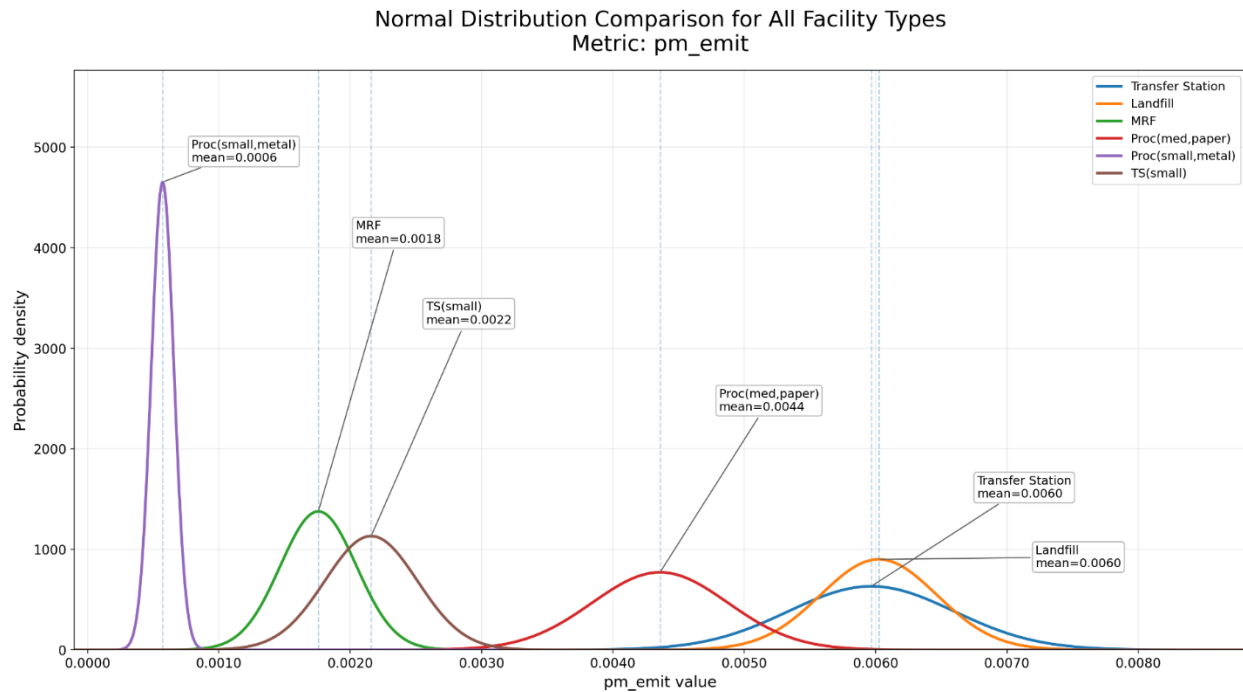


Figure 4.13 On-Site PM Emissions Inventory (Metric Tons)

Compared to NO_x (Figure 4.12) and PM (Figure 4.13), the magnitudes of the GHG are obviously much higher. They also exhibit less variance in the estimate. This is because the GHG factors vary with fleet configuration choice (more feasible combinations for larger fleets) and the horsepower draw whereas NO_x and PM introduce additional variation through the vehicle age draw.

The figures below (Figure 4.15, Figure 4.16, Figure 4.17) illustrate the relative magnitudes of the emissions with a nested pie chart structure. This clarifies that the contribution of off-road heavy vehicle use is very low. Furthermore, the nature of the facilities are such that,

even though landfill and transfer station handle most of the material, their marginal emissions rate is better than other processes.

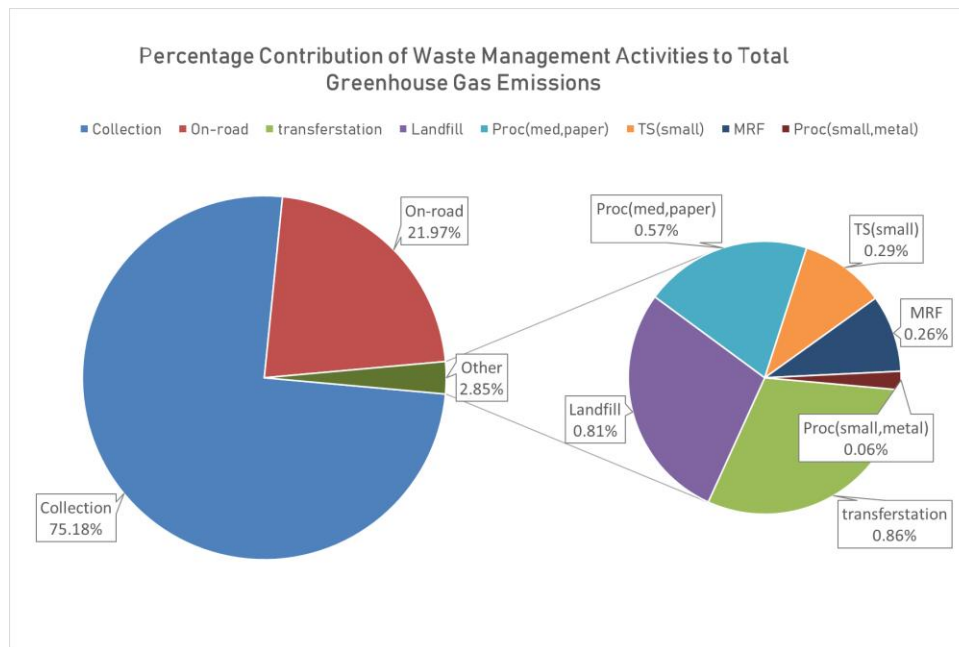


Figure 4.14 Percentage Contribution of Waste Management Activities to Total GHG

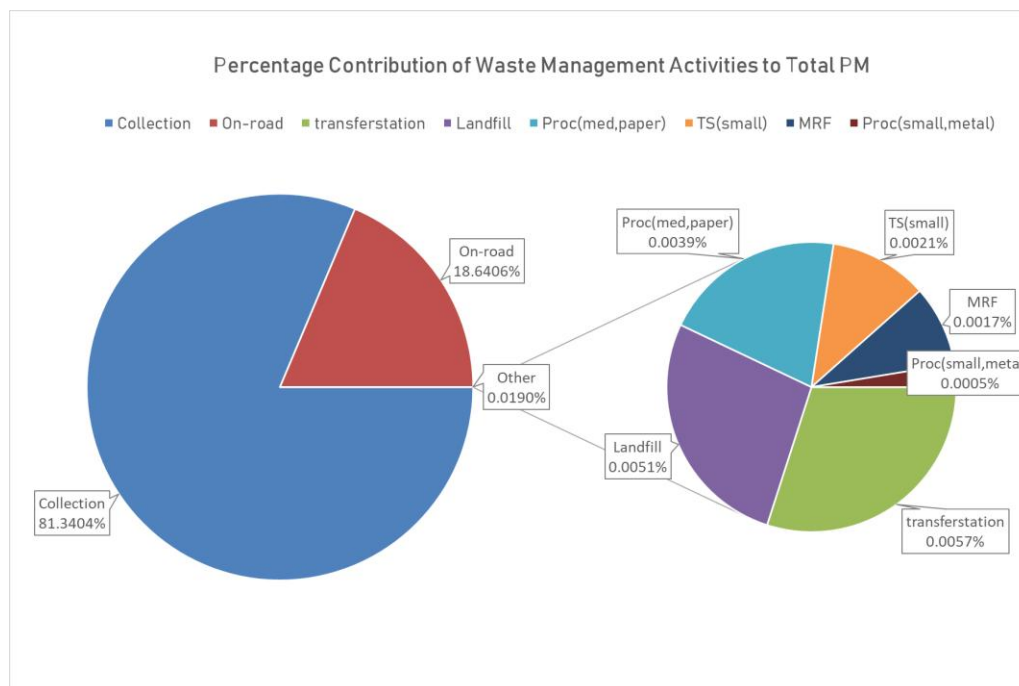


Figure 4.15 Percentage Contribution of Waste Management Activities to Total PM

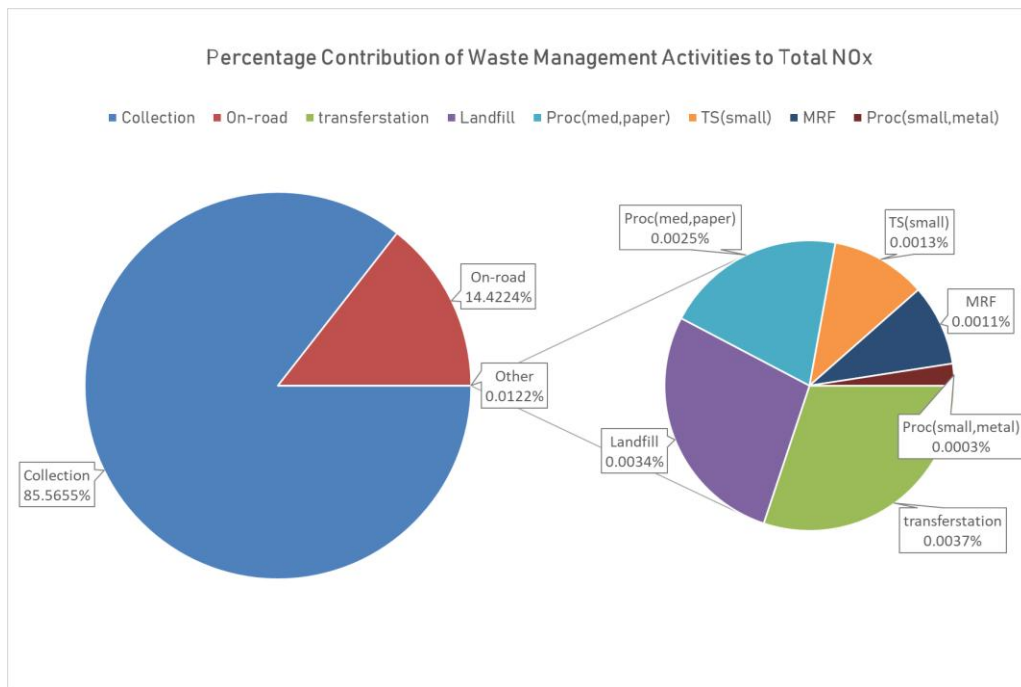


Figure 4.16 Percentage Contribution of Waste Management Activities to Total NO_x

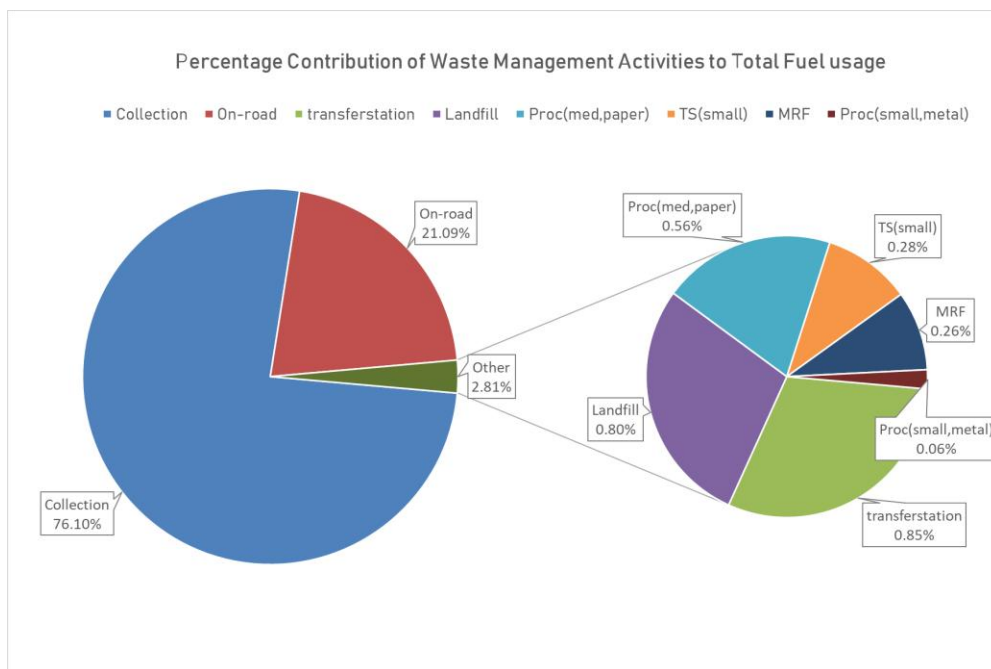


Figure 4.17 Percentage Contribution of Waste Management Activities to Total Fuel Usage

4.4.2 Emissions Inventory of Waste and Recycling Sub-sectors

In this section, each subsystem partition will be assessed for total carbon footprint from all applicable emissions source models identified in §4.3. Care is taken not to double count any quantities nor emissions so that a representative estimate of emissions per ton can be quantified by material, stream, and/or subsector. That being said, on road mileage and emissions are doubled to account for a vehicle's presumed return trip.

First, all Origin Type, Material, and Destination Type combinations are assessed a label as “Aligned” or “Anomalous”. Aligned means that the flow of material should be expected based on the facility types involved (e.g. MSW moving from Transfer Stations to Landfills, Organics moving to Composters, etc.). There are 862 observed combinations assessed (of 5,292 possible). Approximately 88% of the combinations are aligned, but this translates to 99.7% of the total tons and no less than 97.8% of any one material. These high numbers speak to the quality of the facility classification and that the class descriptive labels reflect the reported data well.

Second, every flow is labeled with a disposition. These speak to the nature of the facilities involved and/or material in question. They are: Disposal, Transfer, Diversion, Composting, and Broker Source or Broker Sink. Any materials with destination “Inerts Disposal” or “Landfill” are labeled Disposal. Transfer is made up of any material moving to any class of Transfer Station as well as any material that is *not* Organics moving to any class of Composter. In the latter case, these have been filtered by the alignment to be potentially putrescible materials (Paper, Cardboard, Other, and CD). Aligned composting flows are more narrowly defined as Organics moving to any class of Composter, Anaerobic Digestion, or

Table 4.11 Brokerage quantity by material

Material	Brokered		Not Brokered	
	Quantity	Fraction	Quantity	Fraction
Cardboard	1,287,156	0.87	186,156	0.13
CD	50,238	0.02	2,782,035	0.98
Ferrous	114,007	0.22	404,443	0.78
Glass	69,346	0.11	578,765	0.89
HDPE	26,602	0.43	34,897	0.57
Mixed Metal	17,344	0.07	237,929	0.93
Mixed Plastic	68,763	0.68	31,636	0.32
MSW	41,672	0.00	25,484,697	1.00
NonFerrous	21,900	0.15	120,944	0.85
Organics	1,289,374	0.20	5,215,864	0.80
Other	79,720	0.08	906,202	0.92
Paper	816,105	0.91	81,505	0.09
PET	118,860	0.50	118,778	0.50
SingleStream	24,068	0.02	1,311,716	0.98

Wastewater Treatment facilities. Diversion is made up of any mixed or separated recyclable commodities moving to any class of Processor or Recycler. Broker Sink includes any material moving to any type of Brokerage destination. Flows are similarly defined for all origin types. The only exception is when the origin type is a broker (labeled Broker Source), unless the destination is also a broker, in which case the flow defaults to a disposition of Broker Sink.

In general, flows involving broker endpoints are not actually moving between the locations but more likely represent a changing of ownership. This invalidates the VMT calculations to/from brokers and so they are omitted. Several materials have significant brokerage quantities including Cardboard, Paper, Mixed Plastic, PET, and HDPE. Ferrous and NonFerrous metal, Organics, and Glass have smaller fractions of brokered material while, CD, Mixed Metal, MSW, Single Stream, and Other have less than 10% brokerage. These artificial

Table 4.12 Tons of material per origin, destination facility types

Disposition	Aligned Tons by Destination Facility Type						
_Origin Fac. Type	Landfill	Composter	Transfer Station	Processors/ Recyclers	Other	Inerts	WWTP
disposal	28,200,434					233,702	
TransferStation	26,550,191					228,980	
Composter	612,148					169	
WWTP	482,130						
Processor	378,491					2,430	
Other	169,464						
Landfill	8,009					2,124	
transfer			2,732,469	983,519	352,732		
TransferStation			2,441,357		346,942		
Processor			199,853	972,714	68		
Landfill			55,375				
Composter			27,295				
Other			5,710	10,761	5,721		
WWTP			2,880	44			
composting		3,640,051			111,776		74,880
TransferStation		2,471,788			56,216		19,260
WWTP		551,452			7,261		55,620
Composter		334,578			48,299		
Landfill		227,259					
Other		30,461					
Processor		24,512					
diversion			107,275	931,235	40,140		
TransferStation				841,070	15,884		
Landfill			107,275	69,004	16,256		
Composter				21,159	135		
Processor					7,865		
Grand Total	28,200,434	3,640,051	2,839,744	1,914,754	504,648	233,702	74,880

flows (to brokers) are *not* included in on-road calculations, though the reported outflow may be included in on-site calculations. This marks the largest quantifiable omission of potential on-road material movements and disproportionately impacts recyclables, especially fiber, as indicated in Table 4.11. Upon omission of these artificial flows, the following Table 4.12 of total quantities flowing between each origin-destination facility type combination by disposition results.

The combinations of disposition, associated facility types, and materials are exploited to define mutually exclusive subsectors and pathways, dividing the broader waste and recycling industry into more manageable and interpretable chunks. The query definitions are provided in Appendix B.2. This facilitates the correct application of the emission models as well as the accounting and analysis of statewide emissions and related policy. The ratio of total tonnage and onroad VMT of each reconstituted subsystem is calculated as a statewide average per ton on the 2021 basis, as shown in Table 4.13 and/or Figure 4.18 Miles per ton by material stream.

The material-specific transportation patterns reveal substantial variation in both total tonnage and associated vehicle miles traveled (VMT), reflecting differences in facility geography, processing infrastructure, and routing requirements across waste streams. Municipal solid waste (MSW) disposal constitutes the largest material flow by mass, totaling approximately 25.3 million tons and accounting for nearly 56.0 million VMT. Despite its large scale, MSW disposal exhibits a relatively low transport intensity of 2.2 VMT per ton, suggesting comparatively direct routing and geographically widespread disposal infrastructure. Similarly, construction and demolition (C&D) processing and disposal accounts for approximately 25.9 million tons with 73.1 million VMT (2.8 VMT/ton), indicating somewhat greater transportation demand per unit mass but still relatively efficient movement compared with other streams.

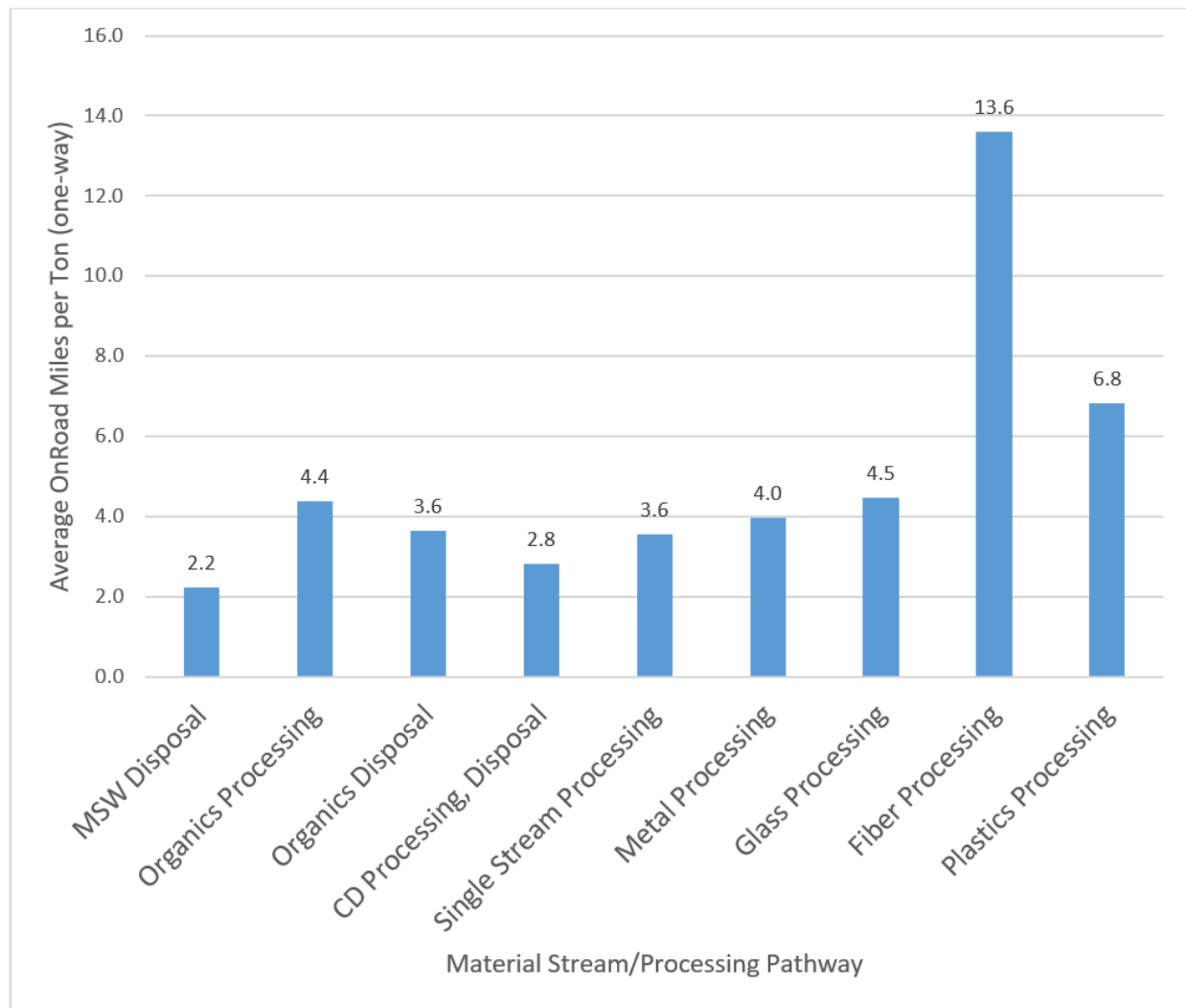


Figure 4.18 Miles per ton by material stream

Table 4.13 VMT per ton by material stream

System	Subsystem Description	Tons	VMT	VMT/Ton
1	MSW Disposal	25,253,551	55,986,884	2.2
2	Organics Processing	4,085,234	17,867,268	4.4
3	Organics Disposal	955,209	3,475,309	3.6
4	CD Processing, Disposal	2,594,102	7,311,039	2.8
5	Single Stream Processing	1,249,446	4,453,784	3.6
6	Metal Processing	741,486	2,938,323	4.0
7	Glass Processing	544,159	2,433,555	4.5
8	Fiber Processing	236,844	3,218,801	13.6
9	Plastics Processing	180,571	1,231,389	6.8

Organic waste exhibits notably higher transport burdens. Organics processing represents more than 4.1 million tons but generates approximately 17.9 million VMT, corresponding to 4.4 VMT per ton, while organics disposal reaches 3.6 VMT per ton. Single-stream recycling processing also demonstrates comparatively high transportation demand, producing roughly 4.45 million VMT from 1.25 million tons (3.6 VMT/ton).

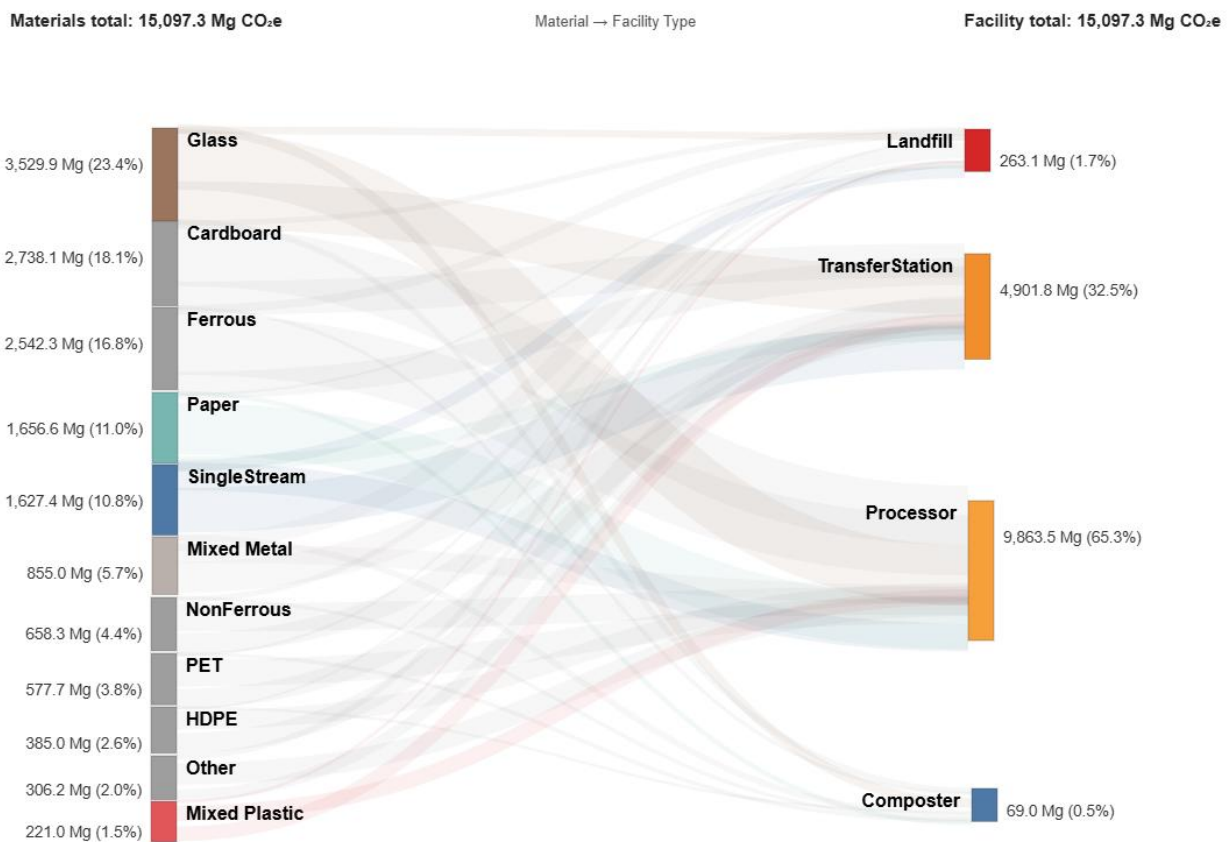


Figure 4.19 Sankey Diagram of Carbon Emissions by Destination

Among individual recyclable material streams, substantial variation is observed, but all are elevated in terms of VMT/ton. Glass processing and metal processing exhibit moderate transportation intensity, at 4.5 and 4.0 VMT per ton, respectively, whereas plastics processing reaches 6.8 VMT per ton, suggesting a more geographically dispersed processing network. Fiber processing displays the highest transport intensity of all categories at 13.6 VMT per ton, despite

relatively modest tonnage (236,844 tons). This finding implies a highly centralized processing structure in which materials may travel long distances to specialized end markets or facilities. Overall, the results demonstrate that transportation demand is not solely a function of material quantity but that it is strongly influenced by the spatial distribution and specialization of downstream waste management and recycling infrastructure.

4.4.2.1 Solid Waste to Disposal

The largest, most identifiable, and consistent subsector represented in the RDRS system is that of municipal solid waste disposal. The majority of this stream consists of one material; municipal solid waste aka MSW comprises 88% of *reported* flows. Another 9% comes from disposed inerts and organics but these fall into separate partitions based on material. These may be highly contaminated or otherwise unsuitable for composting (palm leaves are too fibrous to shred) or reuse (contaminated soil, building materials). The remainder comprises materials which would typically be considered nominally recyclable but are nonetheless disposed for some reason (perhaps industrial waste or otherwise contaminated). These are omitted.

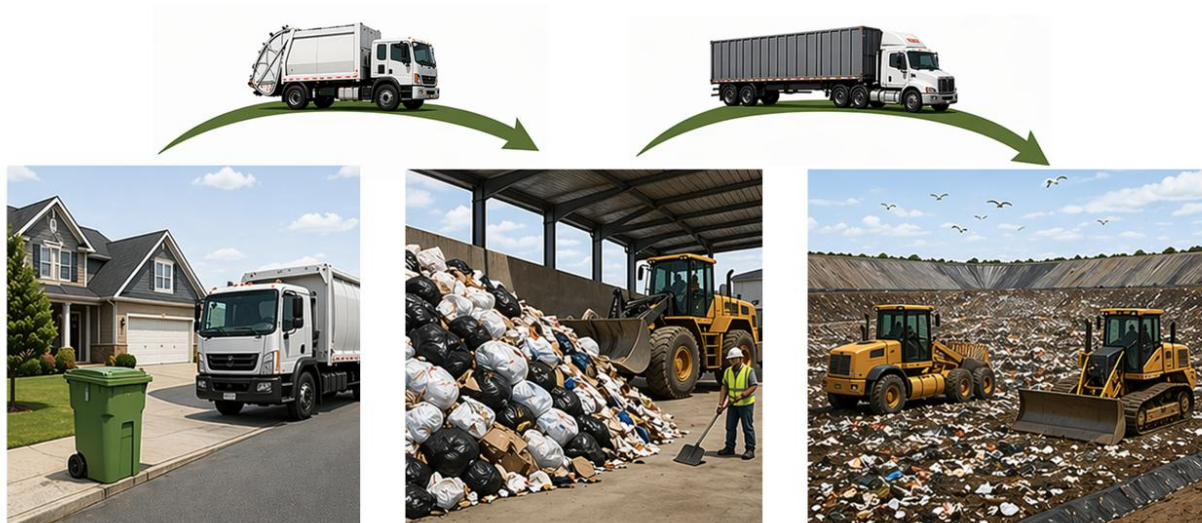


Figure 4.20 Solid Waste to Disposal Path

In general, two on-road legs are involved; collection trucks bring waste from generators to transfer stations and tractor-trailers bring the waste to landfills. Figure 4.20 represents a diagram of this most common situation. In some cases, a small and/or remote transfer station may feed another prior to the waste arriving at a landfill and in other situations, collection trucks attend landfills directly (if the route is close by or at the end of the workday). These variants on path to disposal will be referred to as “multi-step-transfer” and “direct-to-disposal”, respectively.

For collection, the statewide inventory demonstrates the magnitude of transportation activity required to support California's municipal solid waste management system. The modeled residential solid waste collection vehicle (SWCV) fleet traveled approximately **328 million vehicle-miles** during the study year while consuming approximately **96.4 million gallons of fuel**, resulting in an estimated **960,217 Mg CO₂e** of greenhouse gas emissions.

These values illustrate that collection alone represents a major component of the environmental footprint of waste management infrastructure. Because every ton of waste must

first be collected before entering downstream transfer, processing, or disposal pathways, collection activity establishes a substantial baseline level of transportation emissions that exists regardless of subsequent management strategy. Improvements in recycling or diversion therefore do not eliminate transportation impacts, but instead modify the distribution of transportation demand throughout the infrastructure network.

The magnitude of fuel consumption further highlights the scale of heavy-duty vehicle operations required to service California's population. Nearly one hundred million gallons of fuel are consumed annually simply to support residential collection activities, emphasizing that even modest improvements in fleet efficiency or routing could translate into substantial reductions in both fuel use and greenhouse gas emissions.

More broadly, these statewide totals provide an empirical baseline against which the additional transportation and operational impacts of transfer stations, material recovery facilities, processors, and landfills may be evaluated. Rather than viewing emissions as isolated facility characteristics, the results reinforce that system-level emissions emerge from the interaction of infrastructure configuration, transportation demand, and operational activity across the statewide waste management network.

Thus, quantification of emissions is achieved gathering onroad flows which meet the criteria (System 1) and then pulling lists of associated transfer stations and landfills to which on-site models are applied. The total quantity of outflows from transfer station origins are multiplied by the transfer station activity multiplier (1.8 hphr/ton) to generate the emissions at transfer stations whereas the total inflow for landfill destinations are first plugged into equation 4.7 to obtain the multiplier per landfill facility prior to calculating the product. It is important to note

that the “total inflow for landfills” must be patched to include the direct-to-disposal quantities as previously described. These additional 15 million short tons complete the disposal picture.

Table 4.14 Disposal Subsystem Results

Pathway	Emit Source	GHG(MgCO ₂ e)	Percent GHG	Mass (ton)	kgCO ₂ e/ton
Any	Collection	960,217	78.0%	40,230,031	23.9
Via Transfer Station	On Road	159,102	12.9%	25,253,551	9.21
	Off Road (TS)	23,153.7	1.9%		
	Off Road (LF)	50,399.8	4.1%		
Direct-to-Disposal	Off Road (LF)	37,825.9	3.1%	14,976,480	2.53
System 1 Total		1,230,698.3	100.0%	40,230,031	30.6

Table 4.14 indicates the mass of material (MSW) processed and resultant GHG emissions. It is important to take care not to naively double-count material as it exits one facility and enters another. By my estimation, disposal of 40 million short tons of MSW in 2021 contributes 0.27 million metric tons of GHG. Direct-to-Disposal is assessed separately because it is a very significant quantity. In the table it appears to have a superior emission rate on a mass basis but this is because the on road emissions are accounted in the SWCV inventory.

The multi-step transfer leg was not displayed in a separate row because it was such a small amount (less than a quarter million tons per year statewide). The movement *seems* to have an excellent emission rate per ton. This suggests that more short transfers might benefit GHG which of course does not coincide with operational reality (not in this solar system anyway). This is likely because the assumed payload exceeds the daily mass passing through a given facility, resulting in a drastic undercounting of trips. A similar issue can arise in off-road machine hours where hphr demand drops below the level at which a machine is required. Anecdotally, this might correspond to the transfer station in Big Bear, where it is simply a collection of dumpsters where the public drops off their trash for what ultimately equates to commercial/industrial

collection. Any site which generates more than 100 tons per quarter might then just appear as contract hauling – routing and therefore VMT calculations fail because there is no valid geocode endpoint.

A facility-level treatment is necessary to retool very small facilities where the assumptions of payload and heavy machinery requirements forwarded in §4.3 break down. A back-of-the-envelope calculation indicates this problem is pervasive among small transfer stations. Two thirds of small transfer stations (subcategory ‘TS(small)’ in Chapter 3) move less than 5,400 tons (18 tons per day, 300 working days) while 100% fall below minimal heavy vehicle usage thresholds per day, defined as just 4 hours on a single 150 hp loader per day), based on the multiplier 1.8 hphr/ton. Even at 4 hours *per week*, 70% of small transfer stations fall below the threshold ($150\text{hp} \times 0.4 \times 4\text{hr/wk} \times 50\text{week} / 1.8\text{hphr/ton} = 6667 \text{ tons}$) Surveying more transfer stations is necessary to potentially capture economies of scale and find the true thresholds at which operational assumptions fundamentally change.

At this stage of the research, the quantities are small enough to be counted as error swamped by the millions of tons following more common pathways. Omission of the small transfer stations would only result in drops of 2% and 4% in the on- and off-road aggregate emission rates, respectively.

Chapter 5 Conclusions

This dissertation addressed a fundamental challenge in the environmental assessment of waste management systems: the absence of an engineering-based framework capable of estimating emissions when the underlying operational activity is not directly observed. Although regulatory databases provide extensive information on material quantities and facility characteristics, they generally do not report the equipment populations, machine utilization, transportation activity, or operational intensity responsible for fuel consumption and emissions. Consequently, existing inventories frequently rely on generalized emission factors, representative facilities, or simplified assumptions that obscure the substantial heterogeneity of California's waste management infrastructure.

The principal contribution of this research is the development of an empirical framework for estimating operational activity from incomplete observations and integrating those estimates into a unified statewide emissions inventory. Rather than treating collection, transportation, processing, and disposal as independent systems, this work models each component using the most appropriate engineering approach for the available data and synthesizes the results within a common analytical framework. Collection emissions are estimated directly through EMFAC, post-collection transportation is reconstructed through statewide routing and network analysis, and onsite facility emissions are estimated through conditional simulation of equipment populations consistent with observed facility characteristics. Together, these components provide a comprehensive representation of post-collection waste management activity across California.

A central finding of this dissertation is that the greatest challenge in emissions estimation is not the calculation of emissions themselves but the estimation of the activity that generates

them. While emission rates are generally well characterized for both on-road and off-road equipment, the operating hours, fleet compositions, and utilization patterns necessary to apply those rates are rarely available. The conditional simulation methodology developed herein addresses this problem by generating feasible equipment populations subject to operational constraints rather than assuming a single representative fleet. In doing so, uncertainty arising from unknown facility operations becomes an explicit component of the analytical framework rather than an uncontrolled assumption.

The research also demonstrates the importance of viewing waste management as an integrated transportation and infrastructure system rather than a collection of independent facilities. Transportation demand is governed not only by waste generation but also by the organization of transfer stations, processing facilities, and disposal sites throughout the network. Likewise, onsite emissions depend on facility scale, operational practices, and equipment requirements that vary substantially across infrastructure types. Meaningful evaluation of climate impacts therefore requires simultaneous consideration of these interconnected components rather than isolated analyses of individual technologies

Beyond the methodological framework, this dissertation produces a California-specific inventory of post-collection waste management emissions. The inventory provides estimates of greenhouse gases, criteria pollutants, fuel consumption, and equipment activity for multiple subsectors while maintaining consistency across temporal, geographic, and pollutant accounting frameworks. More importantly, it establishes engineering-based generation rates and statewide totals that provide an empirical baseline for future policy analysis, infrastructure planning, and scenario evaluation.

5.1 Principal Findings and Conclusions

The contributions of this dissertation are primarily empirical rather than algorithmic. The objective was not to introduce fundamentally new optimization or machine learning methods, but rather to formulate appropriate engineering models for a poorly characterized infrastructure system, specify defensible assumptions and constraints, and estimate quantities that have not previously been available at statewide scale.

First, this dissertation demonstrates that administrative facility classifications are insufficient for engineering analysis. Facilities sharing identical regulatory labels frequently perform substantially different operational functions, while facilities with different administrative designations may exhibit similar material handling behavior. Consequently, regulatory categories alone do not provide an adequate basis for assigning representative process parameters or emissions characteristics. The empirical facility archetypes developed herein therefore provide a more defensible foundation for engineering analysis and statewide emissions estimation.

Second, this dissertation produces California-specific estimates of transportation activity, facility energy consumption, fuel use, and associated emissions for major components of the post-collection waste management system. Rather than relying exclusively on transferable literature values or stylized assumptions, these estimates are derived from contemporary statewide regulatory data and facility-specific operating characteristics.

Third, the results demonstrate that statewide emissions inventories depend critically on infrastructure configuration. Transportation demand, equipment activity, and processing requirements are functions of facility operations and material routing rather than fixed

assumptions or average distances. Emissions should therefore be viewed as system-level outcomes emerging from the interaction of infrastructure, operations, and material flows.

Finally, this work illustrates a practical methodology for engineering analysis under conditions of incomplete administrative data. By specifying physically motivated models, imposing appropriate engineering constraints, and employing heuristic and simulation approaches where direct identification is not possible, the dissertation provides a reproducible framework for estimating system behavior in complex infrastructure networks where complete operational observations are unavailable.

5.2 Policy Implications for California Waste Management and Climate Planning

The findings of this dissertation suggest several implications for waste management policy and infrastructure planning in California.

First, collection fleet efficiency and decarbonization should remain a high priority. Residential collection alone was estimated to involve approximately 328 million vehicle-miles traveled, consume approximately 96.4 million gallons of fuel, and emit approximately 960,217 Mg CO_{2e} during the study period. Because collection represents a substantial baseline source of transportation emissions, improvements in routing efficiency, vehicle technology, and fleet electrification have the potential to produce significant statewide benefits.

Second, diversion policies should be evaluated using transportation and processing intensity rather than diversion rate alone. Diverting material from disposal to recycling or organics processing changes transportation requirements, facility activity, and energy consumption. Consequently, environmental performance depends not only on the quantity diverted but also on the infrastructure required to process and transport those materials. System-

level evaluation should therefore consider both avoided disposal impacts and additional transportation and processing activity.

Third, emissions inventories and planning efforts would benefit from supplementing administrative facility categories with empirically derived operational archetypes. The analyses presented in this dissertation demonstrate that facilities sharing the same regulatory designation often perform substantially different functions. Incorporating flow-based facility classifications into emissions accounting and infrastructure planning would improve parameter selection, scenario analysis, and policy evaluation.

Fourth, improved reporting of operational activity would substantially reduce uncertainty in statewide emissions inventories. Current reporting systems document material flows and facility throughput but generally do not include equipment inventories, equipment age, annual operating hours, fuel type, electricity consumption, or major processing technologies. Collection of a limited set of standardized operational metrics would permit more direct estimation of emissions and reduce reliance on simulation and engineering inference.

Finally, infrastructure planning should recognize that facility siting and network design are themselves climate policy instruments. Transportation intensity varies considerably across waste streams and depends on the spatial distribution of transfer stations, processors, recyclers, and disposal facilities. Strategic development or expansion of regional processing capacity may reduce long-haul transportation and improve overall system performance. The results therefore suggest that climate impacts are governed not only by technology choices but also by the organization of the statewide waste management network.

5.3 Future Research

Several opportunities remain for future research. Improved reporting of facility equipment inventories and operational activity would reduce reliance on simulation and permit more detailed calibration. Expansion of the framework to organics processing facilities and construction and demolition infrastructure will require dedicated characterization of activity patterns that are presently unavailable. The framework developed herein also provides a baseline for future evaluation of organics diversion and landfill gas policies by integrating transportation, processing, and disposal within a common systems perspective.

Ultimately, the significance of this work lies not merely in producing emissions estimates, but in developing a practical engineering methodology for constructing defensible statewide inventories where direct operational data do not exist. By transforming incomplete administrative records into estimates of physical activity and integrating those estimates across the waste management system, this dissertation provides a foundation for more rigorous environmental assessment, more informed infrastructure planning, and more effective climate policy.

References

Error! Not a valid filename.

Appendix A – Material Dendrogram

A dendrogram is a method of representing hierarchical similarity, as encountered in the hierarchical clustering method. The dendrogram typically has a vertical scale indicating the distance at which they join. While the following figures are not true dendrograms, they do provide a graphical representation of the hierarchy of waste materials complete with pictures. I have divided them into Mixed Waste, Other Waste, and Mixed Recycling. The three

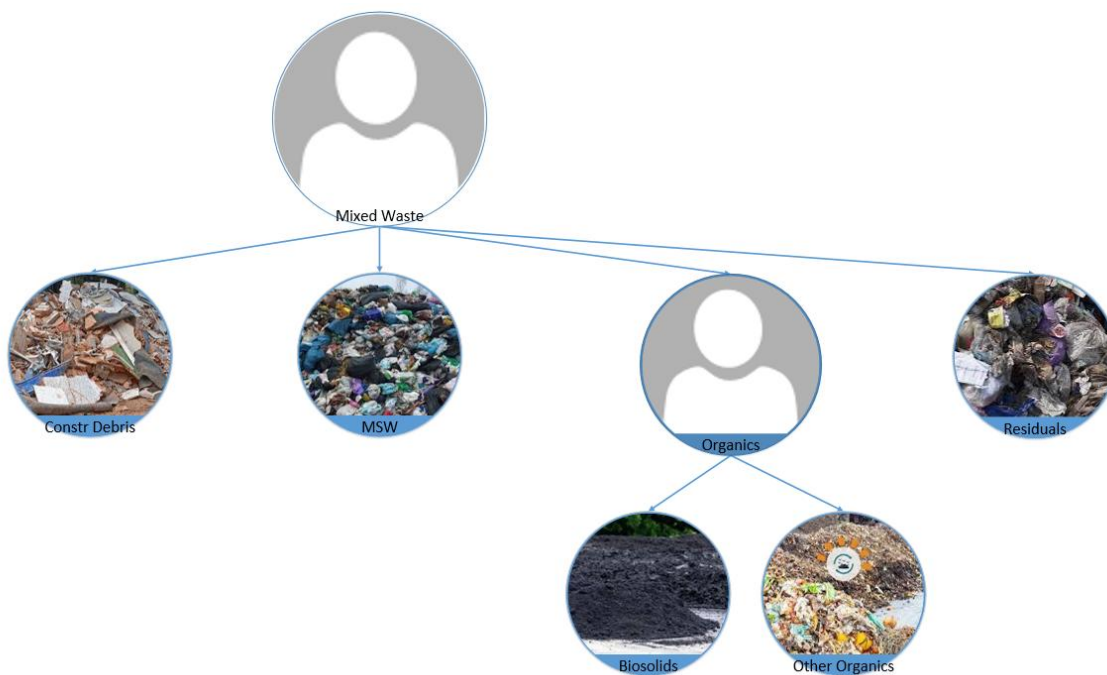


Figure A.0.1 Material Hierarchy (Mixed Waste Branch)

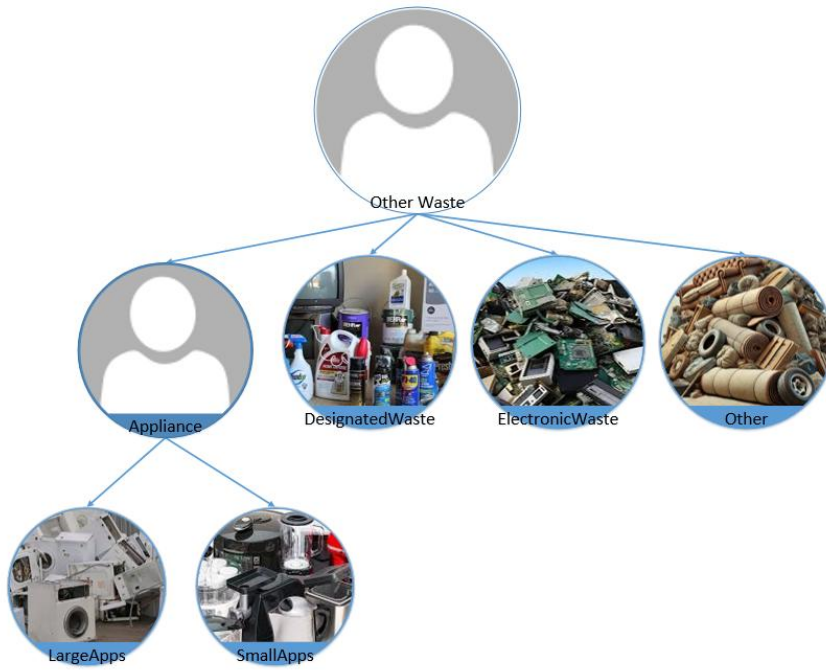


Figure A.0.2 Material Hierarchy (Other Waste Branch)

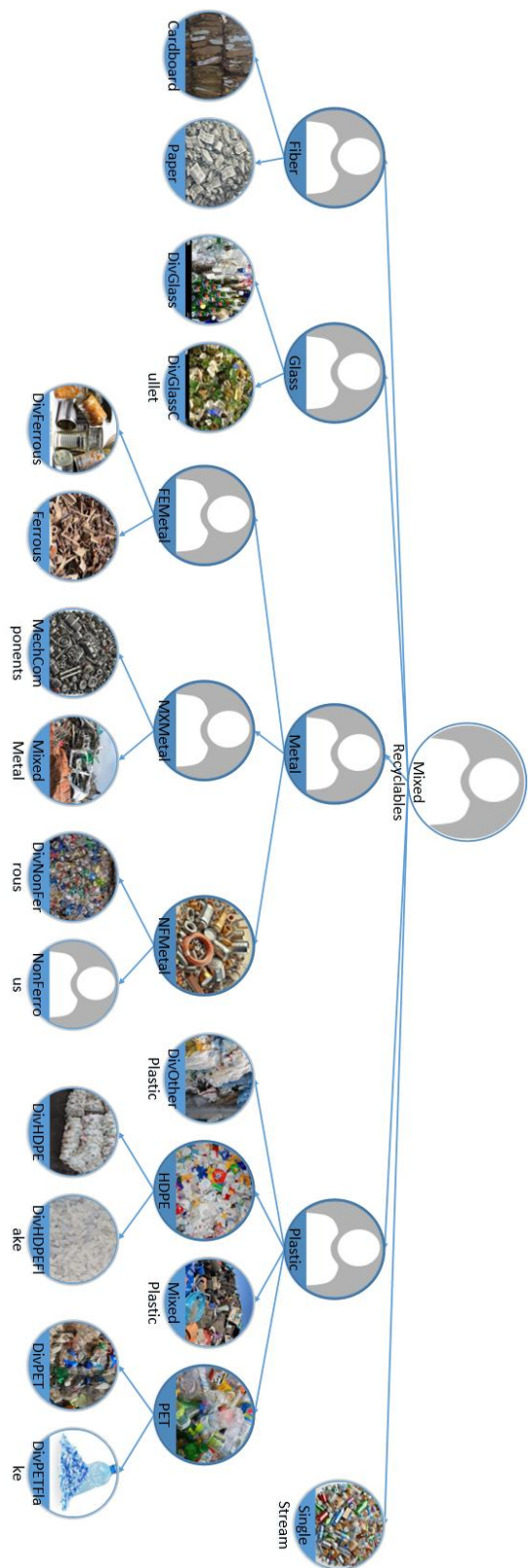


Figure A.0.3 Material Hierarchy (Mixed Recyclables Branch)

Appendix B – Intermediate Calculations

B.1 Survey of vehicle fleet activity

In the absence of comprehensive data regarding heavy machinery fleets and operation time, interviews with managers at several facilities are relied upon to establish average power demands per ton.

The following sites/organizations were contacted and contributed data.

- Irvine Transfer Station (Transfer Station)
- WM Orange (MRF)
- Orange County Waste and Recycling (3 Active Landfills)
- Riverside County Department of Waste Resources (4 Active Landfills)
- Chicago Grade (1 Landfill)

Points of contact were selected based on the number of sites under their jurisdiction and their sizes. Various attempts were made at different numbers to find the “right person” to help, leaving messages, email, and so on. Some organizations were already contacts for tours and other information (e.g. OCWR, WM).

The data request itself involved gathering the annual hours of operation for each machine along with the machines’ characteristics (model, age). I found out that modern CAT machines actually sync with fleet management software that makes this a very simple request for the right person. This cold-calling strategy represents a tiny fraction of what was originally planned, a statewide stratified survey of LCA inputs, which was abandoned based on initial responses. Ultimately, the survey response is just enough to get numbers off the ground for Transfer and MRF operations, while landfills are well-represented by comparison.

Others which were targeted but no response was received include: Inyo County (2 tiny landfills), Recology (multiple sites and types), and Allied (multiple landfills and transfer stations). These would be very valuable to fill in other counties with a variety of sizes and types of facilities with a limited number of contacts. Recology and Allied were chosen particularly because they operate multiple sites but it is clearly harder to penetrate their more corporate structure. On the other hand, the public and smaller private organizations may be more willing to share information. PRA-style requests from Orange and Riverside Counties yielded excellent data on operations at their 7 active landfills sites. And Dannette Fieguth at Chicago Grade was very forthcoming, encouraging my survey efforts and methods, and getting me what I needed within minutes over the phone. Inyo is an exception because I did exchange email with them but my request did not get fulfilled in the end - I think they were not equipped to answer my questions simply because the operations are so tiny and remote.

B.2 System Partitions

The following pseudo-code in Table B.0.1 Partition Scheme defines the conditions used to discriminate subsystem partitions from aligned flows. Each row corresponds to a query defined in terms of disposition, origin facility type (Otype), destination facility type (Dtype), and material. The System column denotes those which are recombined (primarily on the basis of material) to constitute a fuller picture of material movements even across dispositions. Those with a System label of 99 are minor quantities of dubious alignment and/or stand alone as rare cases, possibly idiosyncratic to specific facilities. They are omitted from aggregate onroad VMT/ton calculations but may be included in the relevant on-site emissions models.

Table B.0.1 Partition Schema

System	Query	Tons	Disposition	Otype	Dtype	Material
1	Disposal1	23,985,463	Disposal	TS	LF	MSW
4	Disposal2	1,507,947	Disposal	TS	LF	CD
3	Disposal3	380,018	Disposal	TS	LF	Org
99	Disposal4	167,469	Disposal	TS	LF	<else>
3	Disposal5	575,191	Disposal	Composter or WWTP	LF	Organics
1	Disposal6	973,772	Disposal	<>TS	LF	<else>
4	Disposal7	231,723	Disposal	any	Inerts Disposal	any
2	Composting1	3,602,009	Composting	any	Composter	Organics
2	Composting2	182,926	Composting	any	<>Composter	Organics
99	Composting3	4,839	Composting	any	any	<>Organics
99	Diversion1	37,435	Diversion	any	Other	<>Organics
2	Diversion2	9,735	Diversion	any	any	Organics
9	Diversion3	6,765	Diversion	any	Recycler	Plastics
5	Diversion4	141,014	Diversion	any	Processor	SS, other
5	Diversion5	92,548	Diversion	any	TS	SS, other
7	Diversion6	262,588	Diversion	any	Processor	Glass
6	Diversion7	363,383	Diversion	any	Processor	Metals
9	Diversion8	31,149	Diversion	any	Processor	plastics
8	Diversion9	112,395	Diversion	any	Processor	Fiber
99	Diversion10	5,677	Diversion	any	TS	<>organics, nor Other, nor SS
1	Transfer1	216,141	Transfer	TS	TS	MSW
4	Transfer2	854,432	Transfer	TS	TS	CD
2	Transfer3	290,564	Transfer	TS	TS	Org
5	Transfer4	957,291	Transfer	TS	TS	SS
99	Transfer5	50,407	Transfer	TS	TS	<else>
6	Transfer6	289,856	Transfer	Proc	Proc	Ferrous
7	Transfer7	281,571	Transfer	Proc	Proc	Glass
9	Transfer8	111,921	Transfer	Proc	Proc	Plastics
6	Transfer9	88,247	Transfer	Proc	Proc	Metals (excluding Fe)
8	Transfer10	124,450	Transfer	Proc	Proc	Fiber
5	Transfer11	43,617	Transfer	Proc	Proc	SS, other
9	Transfer12	30,736	Transfer	Proc	Rec	Plastics
99	Transfer13	342,490	Transfer	TS	Transformation	MSW
1	Transfer14	78,174	Transfer	Proc or WWTP or Composter	TS	MSW, other
99	Transfer15	161,509	Transfer	<else>	<else>	<else>

Appendix C – Calibrating Correspondence Analysis

This appendix documents the refinement steps applied to the initial facility sample following an initial correspondence analysis (CA) of the 2021 annual material flows. These refinements were enacted after seeing the structure of the CA-transformed data and its distortion by certain facility subtypes. Each stage of refinement focuses on retaining facilities engaged in transformative or consolidative material recovery and disposal. To enhance the clarity and interpretability of latent dimensions I removed facilities whose material profiles obscure key structural relationships, thereby enhancing interpretability and structure of the result. To be specific, these are removed from CA pattern vector inputs but not from the scope of the study; they are manually labeled per the following descriptions.

Tuning the pattern vector to achieve good feature extraction was analyzed by an iterative procedure. On iteration 0, the full pattern is used to estimate and apply the CA transform.

1. Estimate/Apply CoAn transform under current exclusion criteria
2. Assessment of results for validity, high inertia subgroups
3. If identified:
 - Label
 - Modify exclusion criteria to exclude
 - Return to 1.

C.1 Iteration 0 (n=1686)

Wastewater Treatment Plants (WWTPs) were removed due to their domination of the first two CA dimensions via large biosolid outflows. This is evidenced by the “Out-Biosolids”

column principal coordinate (-3, 2,...) – the rest of them lie along a line, utilizing higher principal dimensions. In-biosolids and Out-residuals indicate that these columns are the most correlated with Out-biosolids (see Figure C.0.1). Review of those upper left quadrant members quickly indicated they were municipal water reclamation facilities. A concerted effort was then made to identify these facilities, based on name conventions and presence of biosolids. Ultimately, 118 were assigned labels and marked for omission from further runs.

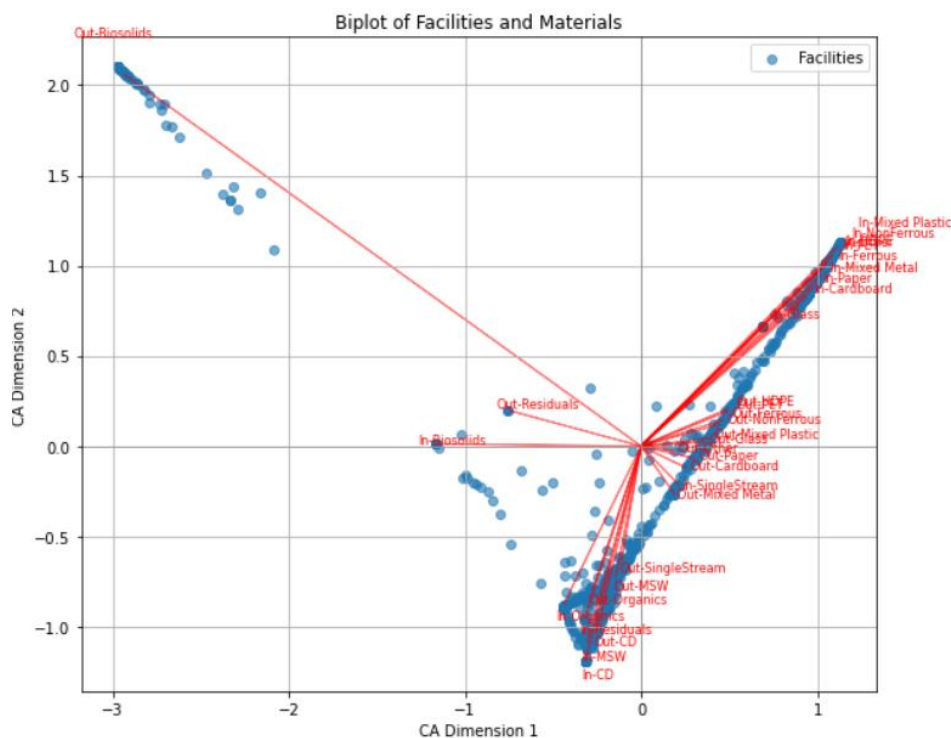


Figure C.0.1 Biplot result for Iteration 0

C.2 Iteration 1 (1568)

Broker/Transporter sinks - facilities acting as terminal endpoints without material transformation—were identified and excluded based on high positive scores on Dimension 1 and broker classifications. These are not true processing facilities and may not even represent a real movement of material but simply a nominal change of ownership.

In the revised CA space (Figure C.0.2), facility positions became more balanced. Dimension 1 captured variation between recycling-oriented inflows and broader mixed material handling, while Dimension 2 separated paper/plastic-specialist inflows from metal-specialist inflows. This dimensional coherence prompted further inspection for remaining distortive influences. Charting the average profiles of the 6 entity types in the database reveals that the high dimension 1 values correspond to Broker/transporters. These are typically not facilities processing waste – at best, they are warehouses or intermediate depots for changes of custody to occur but in many cases are simply a business which markets material. In terms of emissions modeling, they are outside the scope of this research.

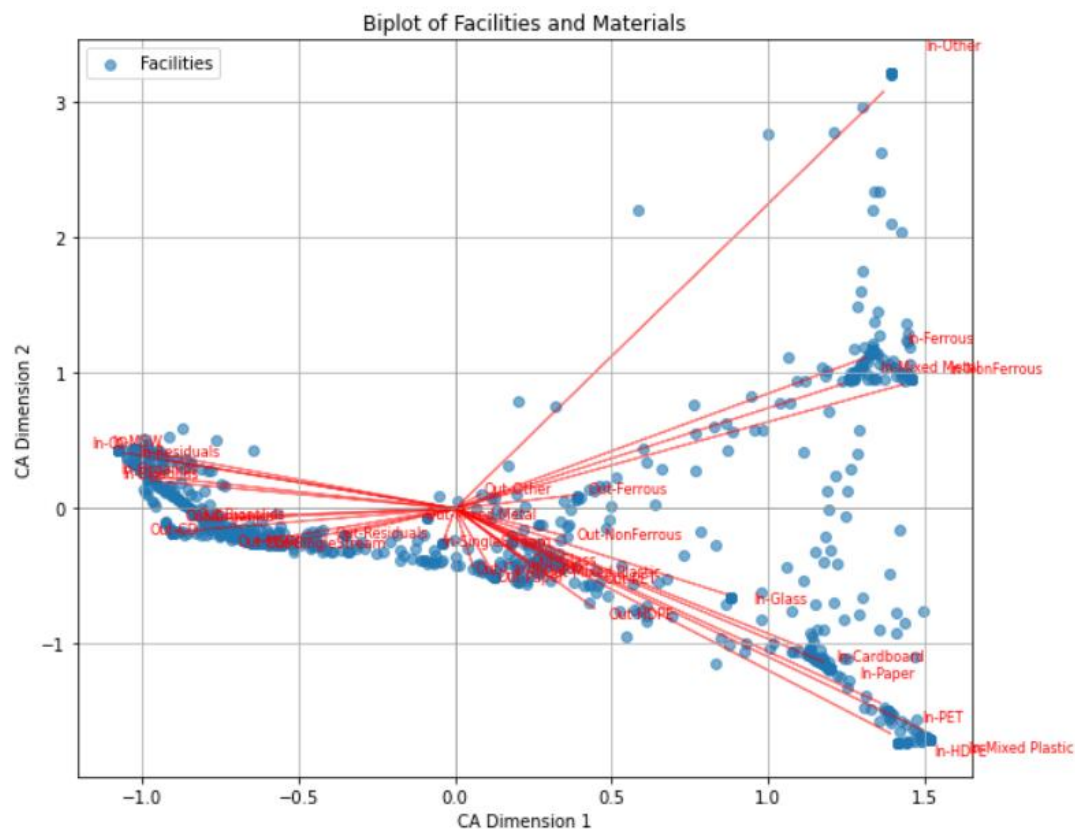


Figure C.0.2 Biplot result for Iteration 1

C.3 Iteration 2 (n=1398)

Bulky material specialists (e.g., tire, mattress, appliance handlers), identified by extreme scores in the first three dimensions. In this iteration, the 2D biplot looks very similar to the previous but with slightly more shaping to the overall cloud. But the 3D biplot clearly reveals that facilities dominated by in-Other flows dominate the inertial spread of the first three dimensions. Its column principal coordinate lies practically orthogonal to the approximate plan formed by the remaining columns. The facilities are identified as predominantly handling bulky goods—tires, mattresses, appliances, and e-waste, reflecting California’s auxiliary diversion programs for these specific materials. These were removed as they are outside the primary research focus.

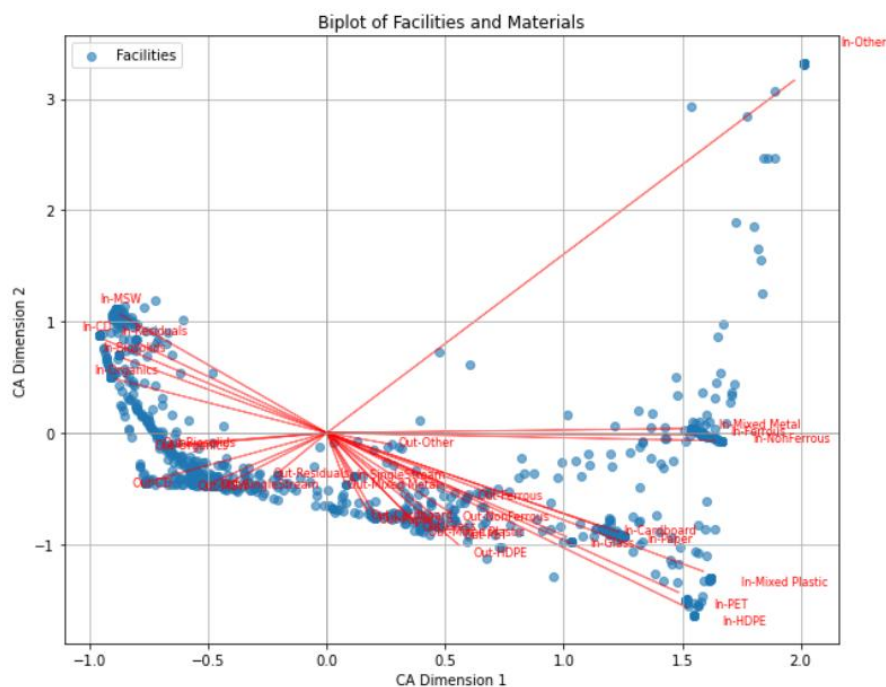


Figure C.0.3 Biplot result for Iteration 2

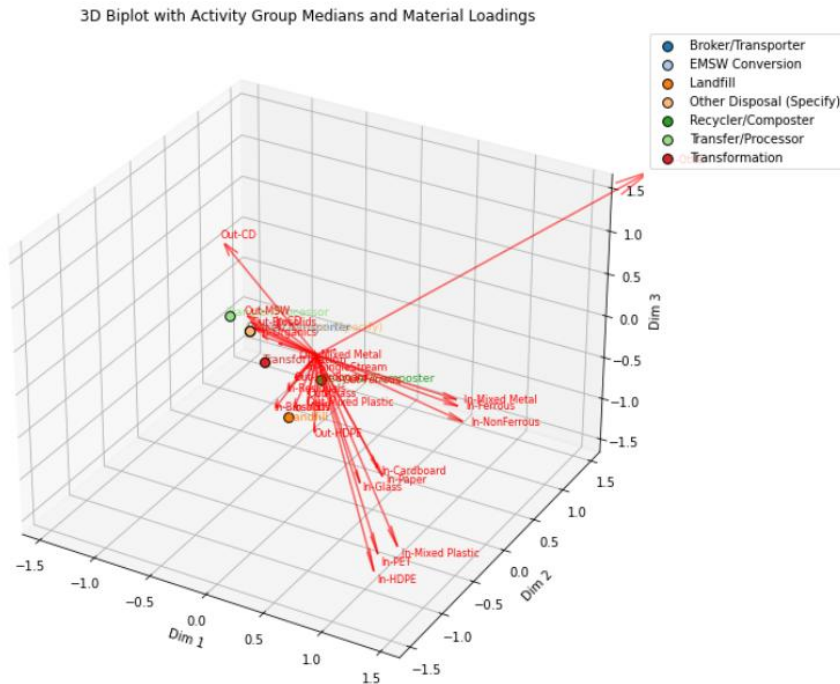


Figure C.0.4 Biplot (three dimensional) for Iteration 2

C.4 Iteration 3 (n=1329)

The CA biplot after three rounds of filtering (Figure C.0.5) shows a well-balanced configuration, with material inflow and outflow vectors distributed across multiple dimensions. Notably, In-SingleStream aligns with Out-Recyclables, while In-Organics and In-CD appear orthogonal, confirming a more interpretable geometric structure for downstream similarity and clustering analysis.

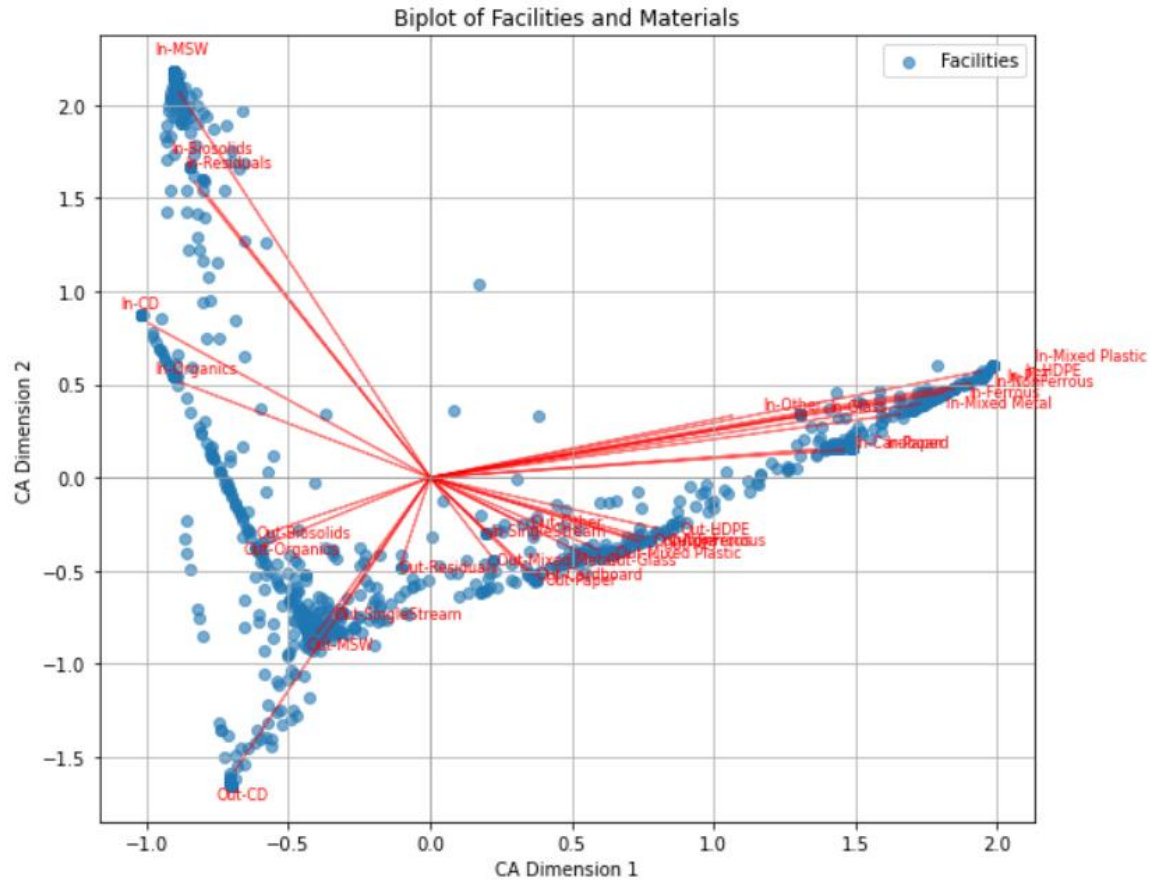


Figure C.0.5 Biplot Result for Iteration 3

The final dataset consists of 1,329 facilities, excluding those whose presence disproportionately influenced the initial latent structure. This cleaned sample serves as the foundation for all subsequent experiments and clustering analyses.

Table C.0.1 Pre-CoAn exclusion conditions

Iteration	Criteria	# Omissions	Facility Description
0	$D1 < -2$	118	Waste Water Treatment
1	$D1 > 0.5$ AND Type="Broker/Transporter"	170	Broker Sinks
2	$D1 > 1$ and $D2 > 0.5$	69	Specialty Diversion (bulky other)

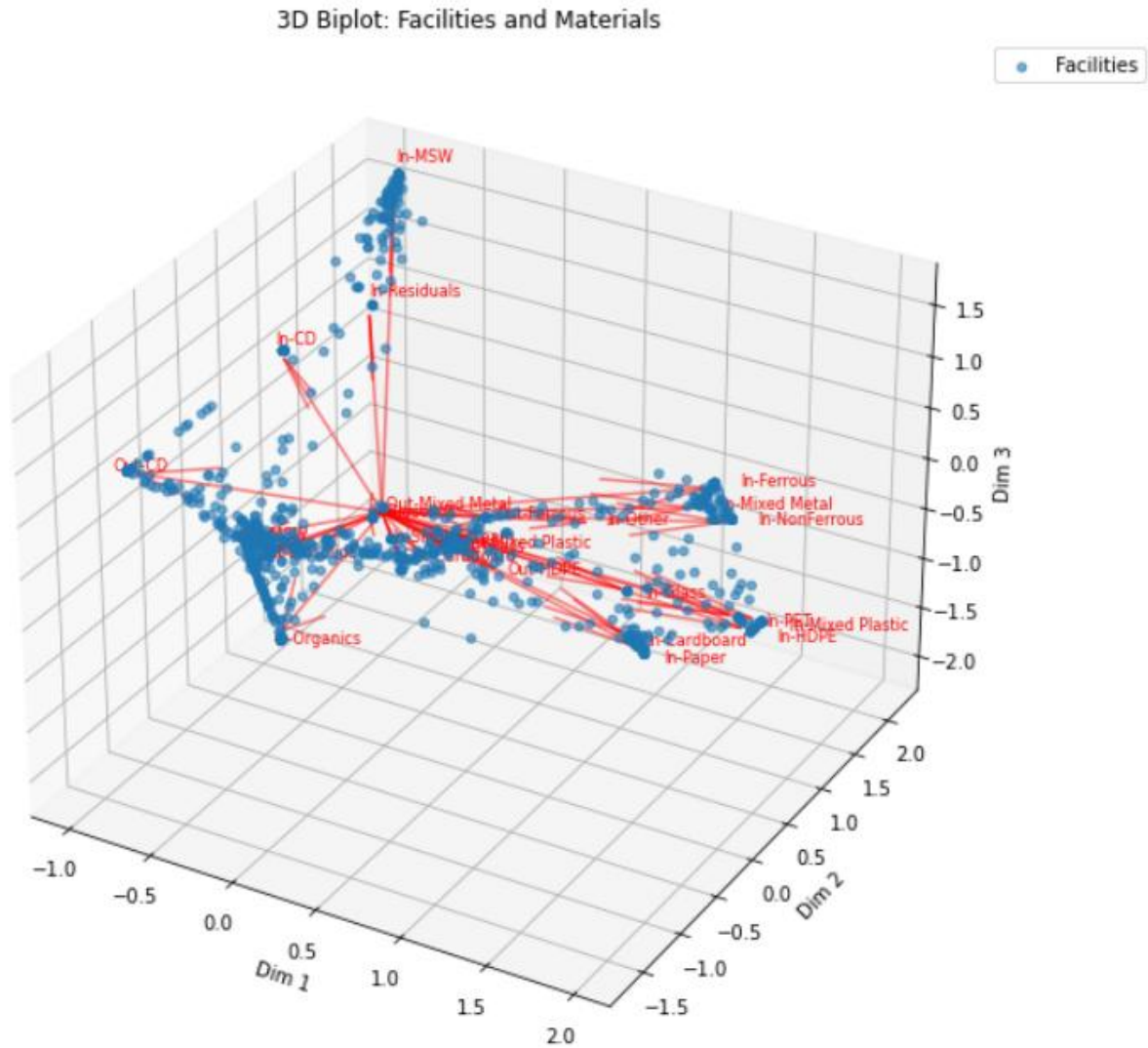


Figure C.0.6 Biplot result (three dimensional) for final iteration

This outcome is very interesting because the In- and Out-CD vectors are in opposition to the In- vectors for recyclables. Also, when viewed in three dimensions, there seem to be natural clustering of the sub-categories of metals, plastics, and fiber, nearby one another in the transformed space but clearly distinct. This suggests that there are some facilities specializing in these classes of materials. In two dimensions, In-CD and In-Organics are virtually on top of one another but in the third dimension they are separated. Most notably, though difficult to discern

due to the compactness and overlapping labels, In-SingleStream is very aligned with Out-vectors for recyclables. These are clumped together in the first few dimensions but would be discriminated by delving into smaller eigen-value-dimensions.

In order to do this, it is necessary to bring in the machine learning to make decisions that I have been making. This is because the display is cumbersome and the thresholds between have become much smaller.

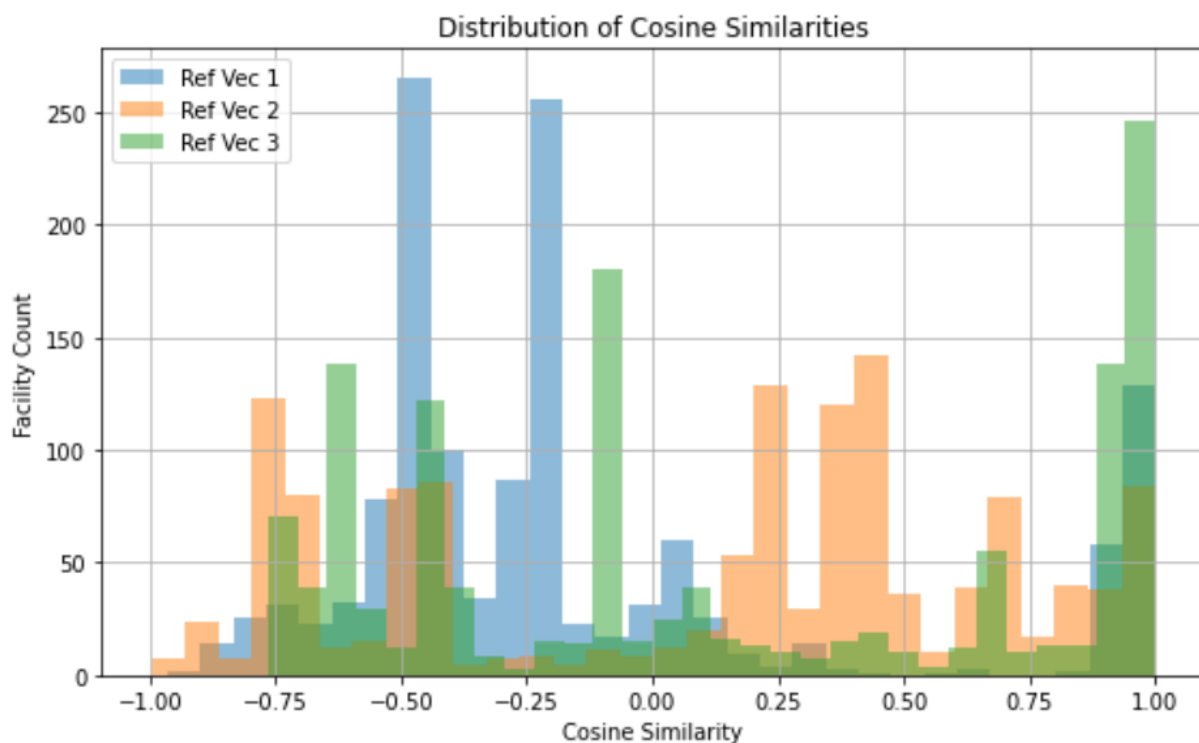


Figure C.0.7 Distribution of Cosines

Examining the histograms of cosine similarity in Figure C.0.7, referenced to three different vectors (landfill, recycler/composter, and transfer/processor, respectively) illustrates that there are distinct clusters of facilities. All three have multimodal peaks throughout the range. Looking at the Ref Vec 1 in blue which corresponds to landfills, we see some 120 facilities with a very high alignment with the average landfill vector (pointing up in the 3D plots). But most of the facilities lie in the -0.5 and -0.25 modes because they are arranged pointing down and right.

C.5 Column Standard Coordinates for the first 17 principal dimensions:

Table C.0.2 Column Standard Coordinates for 17 dimensions

	Dim1	Dim2	Dim3	Dim4	Dim5	Dim6	Dim7	Dim8	Dim9	Dim10	Dim11	Dim12	Dim13	Dim14	Dim15	Dim16	Dim17
Materials																	
In-Biosolids	-0.8498	1.6690	-1.5240	0.2454	1.9423	4.1109	-5.8421	1.5234	-0.3614	-0.7568	0.5822	0.0216	0.1062	0.1176	-0.3279	-0.0639	-0.3589
In-CD	-1.0180	0.8744	0.5913	0.4253	-2.6149	-0.6818	-0.1628	2.7223	-0.7758	0.0416	0.0201	-0.1563	0.0192	-0.0189	0.0058	-0.0041	-0.0006
In-Cardboard	1.3947	0.1589	-0.6623	-1.7965	-0.3905	-1.4443	-1.1153	-0.3172	-0.1745	0.0813	-0.0983	-0.0015	-0.6329	0.4835	-0.2129	0.4910	0.4856
In-Ferrous	1.7685	0.4176	0.8383	2.8002	0.3047	-0.6005	-0.4428	-0.5537	-0.8337	-0.4117	0.2527	-0.0187	-0.4666	2.0006	1.3101	-0.3072	-0.0981
In-Glass	1.3083	0.3417	-0.3741	-0.2408	-0.0816	1.5882	-0.0631	0.8844	1.6963	6.0398	-6.0156	-0.1136	-0.6169	1.0796	1.4005	-1.2193	-0.2001
In-HDPE	1.9479	0.5336	-0.5869	-1.3558	-1.7420	3.3353	2.1113	-0.3434	-0.5392	-1.6098	1.5973	-0.0276	-0.5752	-0.0553	1.4766	-7.7068	1.7517
In-MSW	-0.9061	2.1812	1.6297	-0.7305	0.6316	-0.0648	0.3755	-0.9112	0.2994	0.0260	-0.0295	0.6357	-0.0515	-0.0237	0.0114	-0.0162	-0.0490
In-Mixed Metal	1.6968	0.3585	0.6415	2.1624	0.2384	-0.6416	-0.4734	-0.4360	-0.5738	-0.2620	0.1554	0.0201	-0.4582	0.8225	0.6521	0.0935	-0.1353
In-Mixed Plastic	1.9861	0.6007	-0.5129	-1.0451	-1.6873	2.7916	1.6307	-0.6140	-1.1154	-3.0460	-2.5963	0.0165	0.4728	0.1608	-0.4270	2.7121	-0.6823
In-NonFerrous	1.8489	0.4775	0.4923	1.9211	-0.0521	-0.0507	-0.2361	-0.5650	-0.7127	1.1206	-0.2135	0.1412	1.6374	-3.8230	-3.0192	-0.4390	0.6476
In-Organics	-0.9039	0.5409	-2.0779	0.8113	-0.3914	-0.3819	0.4562	-0.7917	0.6913	-0.0649	0.0097	0.1253	-0.0266	0.2080	-0.3689	-0.1753	-0.4919
In-Other	1.0906	0.3503	0.2038	0.7265	0.1157	-0.6709	-0.3927	-0.2686	-0.1487	0.0948	-0.0084	0.0114	1.5318	-0.7016	0.3993	-0.1840	-0.4157
In-PET	1.8890	0.5013	-0.4568	-0.9185	-1.3196	2.7148	1.5637	-0.0786	0.3130	4.1597	5.3888	-0.0194	0.1021	0.6588	0.3731	2.7047	-0.9273
In-Paper	1.4927	0.1559	-0.7716	-2.2448	-0.4128	-2.0742	-1.5798	-0.4121	-0.1719	-0.1300	0.3327	-0.0238	-0.3215	-1.0048	0.5106	-0.6635	-0.7558
In-Residuals	-0.8046	1.5996	0.7093	-0.3265	0.8049	-0.0370	0.4037	-1.2244	0.0991	0.0052	0.0221	-1.31415	-0.0267	-0.1468	-0.2241	0.0655	0.1791
In-SingleStream	0.1966	-0.3009	0.0199	-0.6244	0.6825	-0.3046	0.2456	0.4336	-0.1844	0.4518	0.0578	0.2161	2.4617	4.7452	-3.5294	-0.2086	3.2227
Out-Biosolids	-0.5718	-0.3053	-0.5038	0.1167	0.4304	0.4015	-0.3313	-0.0765	-0.2869	-0.0326	0.0453	-0.2639	-0.2076	-0.0949	0.0025	-0.0347	-0.4089
Out-CD	-0.7030	-1.6565	1.1079	0.0590	-1.3009	0.4971	-0.9469	-1.2897	0.8383	-0.0429	0.0026	-0.0128	0.0619	0.0563	-0.0835	-0.0239	0.0429
Out-Cardboard	0.3451	-0.5255	0.0654	-0.4035	0.8307	-0.2770	0.3386	1.0678	1.1436	-0.5641	0.2019	-0.0203	1.1721	0.5095	-0.6521	-0.0371	0.4842
Out-Ferrous	0.7782	-0.3412	0.5447	0.8801	0.7519	-0.2085	0.2154	0.9591	1.7935	-1.0420	0.3580	-0.0065	-1.4770	0.2448	0.7308	0.8020	-0.4503
Out-Glass	0.5757	-0.4503	0.2088	-0.0160	0.8989	0.1307	0.6161	1.5843	2.3994	-0.4960	-0.0718	-0.0430	-0.9667	-0.6802	-0.3916	0.1814	0.2219
Out-HDPE	0.8189	-0.2875	0.0076	-0.4307	0.3213	0.9344	1.0839	1.1939	1.6509	-0.9599	0.7257	-0.0423	-0.9829	-0.6008	-0.1371	-2.9075	0.9513
Out-MSW	-0.4061	-0.8799	0.1148	-0.2584	0.8021	0.0687	0.4500	0.2606	-1.1843	0.2002	-0.0469	0.0619	-0.3823	0.0450	-0.3774	-0.2847	-1.1019
Out-Mixed Metal	0.2211	-0.4465	0.4291	0.2397	0.7190	-0.0737	0.3040	0.6204	0.8262	-0.5033	0.1595	0.0194	-0.6306	-0.1399	0.1892	0.2707	-0.4990
Out-Mixed Plastic	0.6080	-0.4079	0.2320	-0.0641	0.8874	0.2756	0.7677	1.7059	2.6764	-1.0108	0.1694	-0.0473	-1.7490	-1.2116	-0.6978	-0.0139	0.2088
Out-NonFerrous	0.7317	-0.3422	0.3559	0.3343	0.8407	0.0372	0.5205	1.4175	-0.7843	0.2090	0.0286	-0.0286	-1.3217	-0.7217	-0.1692	0.4347	-0.0187
Out-Organics	-0.6104	-0.3901	-0.5491	0.0678	0.4894	0.0729	0.2575	-0.0747	1.1303	-0.4376	-0.0225	-0.0365	6.5485	-1.3546	1.3335	0.7372	2.1918
Out-Other	0.3268	-0.2532	-0.0616	-0.0950	0.6477	-0.4989	0.0935	0.5984	1.1303	-0.5215	0.4287	-0.0514	-1.1587	-0.7817	-0.5124	-0.0994	-3.4002
Out-PET	0.3794	-0.3631	0.1974	-0.0155	0.7481	0.2790	0.6988	1.5314	2.4213	-0.5215	0.2977	-0.0129	3.5717	1.0979	-0.6176	0.2932	0.2512
Out-Paper	0.3766	-0.5531	-0.0332	-0.6483	0.8772	-0.5419	0.2065	1.1254	1.1209	-0.6649	0.2947	-0.0129	2.6904	-0.1682	2.1260	0.4181	1.5355
Out-Residuals	-0.1048	-0.4866	-0.0068	-0.1662	0.4242	-0.0817	0.1664	0.3264	0.6714	0.0589	-0.2947	0.0246	2.6904	-0.1682	2.1260	0.4181	1.5355
Out-SingleStream	-0.3144	-0.7257	-0.0499	-0.2987	0.9083	-0.0516	0.4969	0.3913	-0.9579	0.1185	-0.0172	-0.4067	0.6858	0.2801	0.0234	0.0286	0.4629

C.6 Inertias and Eigen Values:

Table C.0.3 Eigen values and contribution to inertia

Dimension	Eigen Value	Inertia	Contribution	Cumulative Contribution
1	0.977	0.955	0.051	0.051
2	0.948	0.899	0.048	0.100
3	0.940	0.883	0.048	0.148
4	0.936	0.876	0.047	0.195
5	0.930	0.865	0.047	0.241
6	0.923	0.852	0.046	0.287
7	0.918	0.843	0.045	0.333
8	0.900	0.810	0.044	0.376
9	0.885	0.783	0.042	0.419
10	0.865	0.748	0.040	0.459
11	0.850	0.723	0.039	0.498
12	0.828	0.686	0.037	0.535
13	0.808	0.653	0.035	0.570
14	0.801	0.642	0.035	0.605
15	0.798	0.637	0.034	0.639
16	0.790	0.625	0.034	0.673
17	0.785	0.616	0.033	0.706
18	0.775	0.601	0.032	0.738
19	0.766	0.587	0.032	0.770
20	0.759	0.576	0.031	0.801
21	0.746	0.557	0.030	0.831
22	0.743	0.552	0.030	0.861
23	0.712	0.507	0.027	0.888
24	0.636	0.404	0.022	0.910
25	0.623	0.388	0.021	0.931
26	0.612	0.375	0.020	0.951
27	0.532	0.283	0.015	0.966
28	0.427	0.183	0.010	0.976
29	0.406	0.165	0.009	0.985
30	0.392	0.154	0.008	0.993
31	0.349	0.122	0.007	1.000
32	0.000	0.000	0.000	1.000

Appendix D – Miscellaneous Methods

Compositional data—such as material flow proportions across categories like paper, plastics, and organics—are constrained by a constant sum, typically 1 or 100%. This introduces unique mathematical challenges. Standard statistical operations like means, variances, or non-parametric tests are invalid in this context because compositional vectors reside in a non-Euclidean space called the simplex. In this space, increases in one component must be offset by decreases in others, meaning components are inherently dependent.

To perform valid statistical analysis on such data, the isometric log-ratio (ILR) transformation is used. This transformation projects compositional vectors from the simplex to a standard Euclidean space where independence and standard geometry apply. Given a composition $\mathbf{x} = [x_1, x_2, \dots, x_D]$, the ILR transform produces a real-valued vector of dimension $D-1$.

Each coordinate in this transformed space is constructed using an orthonormal basis—most commonly generated from a sequential binary partition (SBP) of the components. An SBP is a systematic way of dividing the components of a composition into two non-overlapping groups at each step of the transformation. Each partition defines a contrast between a selected group of parts (numerator group) and the remaining group (denominator group), and the resulting log-ratio forms one coordinate of the ILR-transformed vector.

Mathematically, suppose we begin with a composition of D components. The SBP specifies $D-1$ binary splits. In each step, a contrast is defined between two subsets of parts, labeled $+1$ and -1 , with unselected parts assigned 0. The ILR coordinate for that step is the log of the ratio of geometric means of the two subsets, scaled by a normalization constant:

$$z_k = \sqrt{\frac{r_k * s_k}{r_k + s_k}} * \ln \left(\frac{g(x_i: p_{ki} = +1)}{g(x_i: p_{ki} = -1)} \right)$$

where $p_{ki} \in \{+1, -1, 0\}$ indicates the partitioning of component i at step k , while r_k and s_k are the number of components labeled $+1$ and -1 , respectively.

In implementation, the SBP is typically represented as a D by $D-1$ matrix B , with entries of $+1$, -1 , or 0 . The sequence and structure of partitions encode the interpretive logic of the ILR coordinates. The choice of SBP determines how interpretability is distributed across the ILR dimensions. Though any orthonormal basis is mathematically valid, the SBP offers a structured and interpretable approach tailored to domain knowledge. For example, the first coordinate might contrast organics against all other materials, while the second might compare plastics to fibers, and so on. However, implementation in this research uses default SBP constructs offered by the `composition.ilr` method in python. This matrix is then used for both forward and inverse transformations between the composition and ILR spaces.

D.1 Isometric Log Ratio (ILR) Model

D.1.1 Forward-Transform to ILR space

Prior to transformation, zero values must be handled carefully because the logarithm is undefined at zero. A small positive constant (I utilize 0.0001) is added in place of zeros, followed by renormalization to maintain constant sum.

Once the compositional data has been transformed into ILR coordinates, conventional statistics can be applied. The mean and standard deviation of each ILR coordinate are computed using standard formulas. To test for differences in distributions between groups, such as between clusters of facilities, non-parametric tests like the Kruskal–Wallis test can be applied

independently to each ILR coordinate. This is appropriate because the ILR transformation has rendered the data into a space with independent dimensions.

D.1.2 Back-Transforming Confidence Intervals from ILR Space

When statistics are constructed in ILR space they must be back-transformed to the original compositional space to enable interpretation in terms of the actual material categories.

Furthermore, standard deviations and variances computed in ILR space do not directly map to their counterparts in compositional space. This is because variation in the simplex is not linear, and applying variance formulas after inversion leads to biased results. For this reason, confidence intervals must be utilized to represent variation, constructed in ILR space before inversion.

The ILR transformation is invertible, and its inverse maps Euclidean coordinates back into the simplex. Suppose a confidence interval is calculated in ILR space for a group of samples, yielding endpoints $\mathbf{z}_{low}$ and $\mathbf{z}_{high}$, each of which is a vector of D-1 ILR coordinates. To return to the composition space, the inverse ILR transformation is applied as:

$$\mathbf{x} = C(\exp(\mathbf{B} * \mathbf{z}))$$

In this expression, $\mathbf{B}$ is the matrix denoting the orthonormal basis utilized, and $C()$ represents the closure operation ensuring the resulting vector sums to one: $C(\mathbf{v}) = \mathbf{v} / \sum v_i$

By applying the inverse ILR transformation to both endpoints of the confidence interval, we obtain two compositional vectors: one representing the lower bound and the other the upper bound of the confidence region for each material component. These bounds must be carefully interpreted because the high and low confidence intervals may not display proper ordering of all

dimensions. This is because the high end of the CI for one material necessitates another material to be decreased. As an example assume for a two component estimate that:

$$\mathbf{x}_{\text{low}} = [0.85, 0.15]$$

$$\mathbf{x}_{\text{mean}} = [0.90, 0.10]$$

$$\mathbf{x}_{\text{high}} = [0.99, 0.01]$$

This suggests that the confidence interval for component two is in reverse order where the “lower” bound is higher and “higher” bound is lower than the mean. However, the interval itself, properly ordered, needs to be interpreted. We are 95% confident that the mean of the first component lies between 0.85 and 0.99, and that the mean of the second component lies between 0.01 and 0.15 (as it is bound by the unit sum constraint of compositions). As suggested by the example numbers, the CI ranges are generally asymmetric about the mean once inverted back to the original simplex space.

D.2 MRF Material Flow Model

$$X_s = X = \text{reported quantity of inflow } s$$

$$Y_i = \text{reported quantity of commodity outflow } i$$

$$p_i^{(y)} = \text{purity of target commodity in commodity outflow } i$$

$$r_i = \text{recovery rate of commodity } i \text{ to target commodity outflow } Y_i$$

$$c_i^{(s)} = \text{fraction of material } i \text{ in inflow stream } s$$

$$\tilde{X}_i = \text{unreported quantity of commodity inflow } i$$

$$p_i^{(x)} = \text{purity of target commodity in commodity inflow } i$$

$X \rightarrow \text{MRF} \rightarrow Y_i$

A commodity is defined as any material for which the MRF is targeting recovery. These exit the facility as “commodity outflows”. One or more mixed stream(s) with a characteristic composition of commodities and refuse enters the facility. Each commodity to be separated (including refuse) is concentrated into streams which are named for that target material. The fraction of the total amount of commodity entering the system that is correctly sorted into its target outflow stream is the recovery rate. The complement to that rate is the fraction of that commodity ending up in the wrong commodity outflow stream.

Formulation 0: Perfect Recovery from Single Inflow to Pure Commodity Outflows

For this initial example, purity and recovery are assumed to be 1 for all commodity outflows (including the refuse stream). $\mathbf{X}$ indicates the total inflow of a mixed waste stream which consists of $i+1$ parts where i is the number of commodities targeted for separation. The ‘+1’ is for the contamination fraction denoted “Refuse”. The fraction of commodity i in $\mathbf{X}$ is denoted c_i and the contamination rate is c_{Refuse} . These are constrained to sum to 1.

$$c_i = X_i/\mathbf{X} \quad \text{AND} \quad c_{\text{Refuse}} = X_{\text{Refuse}}/\mathbf{X} \quad (\text{D.2.1})$$

$$\sum_i c_i + c_{\text{Refuse}} = 1 \quad (\text{D.2.2})$$

Y indicates the total outflow which is also divided into $i+1$ parts, except these parts are represented as separate streams, not as a single, mixed stream. Mass balance is enforced.

$$X = Y \quad (D.2.3)$$

When purity and recovery are unity, the result (equation 4) is trivial and essentially the definition of concentration.

$$Y_i = X_i = Xc_i \quad AND \quad Y_{Refuse} = X_{Refuse} = Xc_{Refuse} \quad (D.2.4)$$

Formulation 1: Single Inflow to Pure Commodity Outflows

If recovery rates are less than 1, some commodity material is lost in the refuse stream. This is distinct from purity, which measures that fraction of an outflow stream that is the target material. For this example, purity is assumed to be 1 for all commodity outflows (excluding the refuse stream). This implies that inflow contamination is perfectly rejected to the refuse outflow stream (typically referred to as “residual”) but not necessarily that the recovery rates for commodities are 1.

Continuing from the first example, equations [1], [2], and [3] and related definitions still apply. Additionally, the recovery rate r_i is defined as the amount of a commodity which is correctly sorted into the corresponding commodity outflow stream Y_i , relative to the total entering the system X_i . Equation [4] is replaced by equations [5] and [6].

$$r_i = Y_i/X_i = Y_i/Xc_i \quad r_i \in (0,1) \quad (D.2.5)$$

The amount of commodity i incorrectly rejected from the corresponding commodity outflow stream is the quantity $X_i - Y_i$. Note that mathematically, the assumption of pure commodity

streams implies that refuse has a recovery rate of 1 and that rejected commodity ends up in the refuse stream only.

$$Y_{Refuse} = Xc_{Refuse} + \sum_i (X_i - Y_i) \quad (D.2.6)$$

Substituting equations [5] and [2]:

$$Y_{Refuse} = X[1 - \sum_i c_i r_i] \quad (D.2.7)$$

If Y_{Refuse} , c_i , and r_i are known, X may be estimated as:

$$X = \frac{Y_{Refuse}}{1 - \sum_i c_i r_i} \quad (D.2.8)$$

This quantity adjusts for unrecovered commodities lost in the refuse stream.

PURITY ≤ 1

When purity of the commodity streams are less than 1, equations [1] through [3] and related definitions still apply. The definition of recovery remains the same conceptually, but it is important to distinguish at this point that Y_i is no longer 100% commodity i, invalidating equation [5]. First, defining purity of stream i as the fraction of Y_i consisting of the recovered portion of inflow X_i we obtain equation [7].

$$p_i = Xc_i r_i / Y_i \quad p_i, r_i \in (0,1) \quad (D.2.9)$$

This begs the question, what materials make up the contaminated fraction of Y_i ? For simplicity, we will not distinguish other components beyond contamination with refuse. The total amount of refuse constituting the impurity in commodity streams must equal the entering quantity of refuse (Xc_{Refuse}) that is not rejected to the refuse stream ($1 - r_{Refuse}$). Thus, equation 8 must hold:

$$\frac{\sum_i Y_i(1-p_i)}{c_{Refuse}(1-r_{Refuse})} = \mathbf{X} \quad (\text{D.2.10})$$

Where r_{Refuse} is the fraction of incoming refuse correctly sorted into the solid waste for disposal stream. Given observed Y_i , estimates of outflow purity, inflow contamination c_{Refuse} , and refuse “recovery” rate r_{Refuse} , the inflow quantity $\mathbf{X}$ can be estimated by equation [8] subject to constraints that $Y_i \leq \mathbf{X}c_i$.

Appendix E – Emissions Modeling and EMFAC

EMFAC, developed and maintained by the California Air Resources Board, is California’s regulatory emissions model for mobile sources. It provides internally consistent estimates of emissions by region, calendar year, vehicle or equipment class, fuel, and pollutant, and is used statewide for planning, regulatory analysis, and policy evaluation (CARB, 2024). Conceptually, EMFAC outputs fall into two distinct categories. Inventories represent complete, internally modeled emissions totals for defined fleets, derived from CARB’s assumptions regarding vehicle populations, age distributions, activity levels, and operating conditions. Rate libraries, by contrast, provide marginal emission factors (e.g., grams per mile or grams per horsepower-hour) that can be applied externally to user-supplied activity estimates. While inventories are well suited to sectors with clearly defined fleets and standardized duty cycles, rate libraries are more flexible and can be coupled to empirical or custom-modeled activity data where system-specific inventories do not exist. That is, the fleet and activity profile is built from the ground up.

A second critical distinction is between on-road and off-road modeling frameworks. On-road vehicles are characterized by travel on public roadways and are modeled in EMFAC on a per-mile basis, with rates conditioned on speed, vehicle vocation, fuel type, and region. Off-road equipment, by contrast, operates primarily on site and is modeled on a power-based metric (horsepower-hours), with emissions dependent on rated horsepower, load factor, regulatory tier, age, and cumulative usage. These methodological differences reflect both physical operating realities and long-standing regulatory practice, and they necessitate different forms of activity estimation and data integration.

EMFAC divides California into CoAbDis subareas, so-called because the boundaries are determined by the intersection of county lines, air basins, and management districts. There are 15 basins, intersecting CA's 58 counties – counties are typically contained fully but may be split by one or more basin boundaries. Riverside County is exceptional because it is split by the South Coast and Mojave Desert basins as well as 2 separate management districts. Ultimately there are 69 CoAbDis subareas, displayed in Figure E.0.1 Map of CoAbDis Subareas.



Figure E.0.1 Map of CoAbDis Subareas

E.1 On-Road Model implementation

E.1.1 truck type, fuel type

EMFAC organizes heavy-duty vehicles into source types defined primarily by gross vehicle weight rating (GVWR) and fuel, with some additional distinctions by trip length or body type, but it does not explicitly encode the full diversity of vocational operations within each category. Yet regulatory updates and supporting research consistently show that parameters such as average speed distributions, idling fractions, start frequency, soak times, and braking intensity vary sharply across vocations and regions. This may limit the overlap between EMFAC's generalized activity patterns and the strongly vocation-specific duty cycles observed in real fleets.

Closely related to the vehicle types are the engine fuels employed. Emissions rates vary by fuel type (e.g. combustion of natural gas does not release oxides of sulfur and electric power does not release tailpipe emissions). Each type of fuel has a different fraction of estimated vehicle miles in each subarea. I utilize a miles-weighted average emission rate for each species by sub-area. The fleet fuel makeup is one of the largest contributors to changes in forecasted emissions because forecasts of future years (calendar year) have fewer diesel vehicles and more alternative fuels. To aggregate across fuel types, the miles-weighted average emission rates for each pollutant are calculated. This is relevant to my modeling because I am merging accounting for different fuel distributions in different vehicle types.

Table E.0.1 Statewide Vehicle Type 7 (GVWR8) and Fuel Distribution, Calendar Year 2021

<i>Vehicle Type</i>	<i>Percent Using Diesel Fuel</i>	<i>Percent Using Natural Gas Fuel</i>	<i>Statewide Vehicle Population (1000s)</i>	<i>Percent of Statewide Vehicle Population</i>
<i>Port Drayage</i>	96.90%	3.10%	21.7	7.30%
<i>CA-registered, goes Out-of-state</i>	99.85%	0.15%	49.3	16.70%
<i>Out-of-state registered, comes to CA</i>	100.00%	0.00%	63.2	21.40%
<i>Single Dump</i>	95.36%	4.64%	13.6	4.60%
<i>Single Other</i>	95.41%	4.59%	28.9	9.80%
<i>Solid Waste Collection</i>	48.20%	51.80%	16.3	5.50%
<i>Tractor</i>	97.19%	2.81%	70.8	23.90%
<i>Other (Public, Utility, Concrete Mix)</i>	93.80%	6.20%	32.2	10.90%
<i>TOTAL</i>	94.90%	5.10%	296	100.00%

E.1.2 Speed distribution (in)variance

The distribution of mileage in each bin, used for the calculation of inventories, is uniform for most of the subareas - exceptions are in the Los Angeles area (LA, SB, Riverside, Orange, Ventura, and Imperial County have unique speed distributions). Speed is divided into 5 mph bins ranging from 5 to 85 mph. The impact of speed on emissions is significant in that some pollutants' rates of release are non-monotonic. Most see the highest emission rate at low speed, a steady decrease to a minimum rate at about 50-60 mph and then a slight uptick at higher speeds. The exception is tailpipe particulate matter which increases with speed.

Some examples of EMFAC-based miles-weighted speed probability density function are shown in Figure E.0.2below. The drayage categories of Oakland, LA and Other Class 8 Trucks have the same distribution, as do the categories marked "single". Tractors and Single trucks have a large density of speeds in the 55 to 70 mph range. Port does as well but it has a much stronger

shoulder from the 20 to 50 mph range. Solid waste collection vehicles have a very diffuse shape with a nearly uniform distribution of speeds from 15 to 65 mph.



Figure E.0.2 Default Speed Distributions for T7 Vehicle Categories

Analysis of speed profile variation by vehicle fuel was performed. For each combination of zone $\times$ year $\times$ class, a χ^2 test of homogeneity was applied to the contingency table of VMT by speed bin and fuel type.

- The null hypothesis H_0 states that the underlying distribution of VMT across speed bins is identical across fuels.
- The test statistic is computed from observed versus expected frequencies across all bins and fuels, with degrees of freedom equal to $(B - 1)(F - 1)$.

Every χ^2 test returned $p > 0.05$, confirming that speed-bin proportions of VMT for Diesel, Natural Gas, and Electric vehicles are indistinguishable within each vehicle category and region. Consequently, fuel-specific speed weighting is unnecessary for subsequent emission-rate aggregation.

Similar tests were completed across vehicle classes (within each zone) and across zones (within each vehicle class). Some zones exhibit statistically significant ($p < 0.05$) speed differences by class while only one class (T7 Single Other Class 8) indicates invariant speed profile across zones. This analysis supports treating the speed distribution implied by EMFAC's mileage factors as fuel-invariant (for a given vehicle class and zone) and indicates that class and zone level speed profile variation remains relevant.

These findings justify postponing fuel-type aggregation until after disaggregate emission rate lookups based on speed, vehicle class, and zone. So, the fraction of vehicle class's zonal VMT (V) utilizing a certain fuel type (t) is calculated, and these weights (f) are used to create a composite emissions profile reflecting the given fuel-share mix.

$$f(c, z) = \left\{ \frac{V_{t,c,z}}{\sum_t V_{t,c,z}}, \dots \forall t \right\} \quad (\text{E.1})$$

While an elegant matrix formulation is achievable, I have not gone so far as to formalize these into the graph database. I can conceive of a traversal which accesses model components, parameters and information on how to apply them or otherwise configure them as inputs to deterministic emissions calculations.