第 3 章　分析データおよび分析方法

3.1　本章の目的

本章では、分析データの元となる家庭 CO_2 統計[2)]の調査票情報および気象データの概要を示した上で、これらのデータに基づく特徴量データの作成および分析対象世帯の選定についてその方法と結果を示す。また、第 5 章および第 6 章において採用した各種分析方法の概要を述べる。最後に、分析に使用したプログラミング言語環境およびライブラリ・パッケージの一覧を示す。

3.2　分析データ

3.2.1　家庭 CO_2 統計の調査票情報

本分析で用いるデータは、家庭 CO_2 統計の調査票情報であり、統計法第 33 条に基づいて二次的利用の申請を行い、環境省より提供を受けた。

家庭 CO_2 統計は、全国の世帯を対象とした政府の一般統計調査であり、家庭からの二酸化炭素排出量やエネルギー消費量の実態を把握することを目的としている。同調査は、平成 29 年度（2017 年度）に開始以降、毎年度※、全国 13,000 世帯を対象に実施され、電気・ガス・灯油の月別使用量や太陽光発電の発電量・売電量の他、住宅の仕様や機器・設備の保有状況・使い方等についても聴取している（図 3.1）。

本分析では、2025 年 3 月時点で入手可能であった、6 カ年度分（平成 29 年度～令和 4 年度）の調査票情報を用いた。

※ 隔年実施に変更され、令和 6 年度以降の偶数年度は調査を実施していない（2026 年 3 月末日 時点）

環境省
Ministry of the Environment

本文へ　English　キーワード検索　検索　ヘルプ

ホーム　環境省について　政策　法令　報道・広報　白書・統計　申請・手続き

地球環境・国際環境協力

ホーム ＞ 政策 ＞ 政策分野一覧 ＞ 地球環境・国際環境協力 ＞ 地球温暖化対策 ＞ 家庭部門のCO2排出実態統計調査（家庭CO2統計）

家庭部門のCO2排出実態統計調査（家庭CO2統計）

１．調査の概要と調査票

環境省では、家庭からの二酸化炭素（CO_2）排出量やエネルギー消費量の実態を把握するため、全国の世帯を対象に政府の一般統計調査として「家庭部門のCO_2排出実態統計調査」（家庭CO_2統計）を実施しています。
本調査は、統計法（平成19年法律第53号）に基づく政府の一般統計調査として平成29年度から実施しています。

３．調査の内容

（1）調査事項

調査項目	主な調査内容
①月別のCO_2排出量を把握するためのエネルギー使用量等について	電気、ガス、灯油、ガソリン、軽油
②太陽光発電について	月別の発電量、売電量、太陽電池の総容量
③設備・機器について	HEMSの有無、家庭用蓄電システムの有無、家庭用コージェネレーションシステムの有無・売電契約有無
④世帯について	世帯員、平日昼間の在宅者、世帯年収
⑤住宅について	建て方、建築時期、所有関係、延床面積、居室数、二重サッシ・複層ガラスの窓の有無
⑥家電製品等について	テレビ・冷蔵庫・エアコン・照明等の使用状況、家電製品に関する省エネ行動、使用場所ごとの照明種類、照明に関する省エネ行動
⑦給湯について	給湯器の種類、冬と夏の入浴状況、入浴やお湯の使用に関わる省エネ行動
⑧コンロ・調理について	コンロの種類、用意する食事の数、調理に関する省エネ行動
⑨車両について	自動車等の使用状況、燃料の種類、排気量、実燃費、使用頻度、年間走行距離、交通手段の利用状況、自動車に関する省エネ行動
⑩暖房機器について	保有状況、使用状況
⑪省エネルギー行動について	実施理由

※調査事項は最新年度の調査のものです。過去の調査事項は、各年度の調査の概要を参照ください。

（2）調査の時期
調査実施年の４月から翌年３月までの毎月（12か月間）実施します（主にエネルギー使用量等）。このほか、夏季調査と冬季調査をそれぞれ実施し、合計14回分の御回答をいただきます。

（3）集計結果の公表
・公表の方法
速報値では、電気のCO_2排出係数が前年度値となっています。事業者ごとの電気のCO_2排出係数の公表後に、再計算を行い、確報値を公表します。
・公表の時期
速報値は調査実施年の翌年10月までに、確報値は調査実施年の翌々年３月までに公表予定です。

（4）調査の地域及び対象
・地域：全国　※詳細は「１．調査の概要と調査票」を御参照ください。
・対象：店舗併用住宅等を除く世帯13,000　（母集団数：約50,000,000）
※10地方３都市階級の30区分に層化しています。地方区分については、エネルギー消費の地域特性を踏まえ、また、国勢調査や家計調査等の既存統計調査の区分を参考に10区分としています。また、インターネット調査モニターからの選定による調査では、報告者の都市部への偏りが懸念されるため、都市階級での層設定を行っています。具体的には、都市階級別住宅に住む主世帯数（国勢調査）による比例配分を行っています。
・対象者の抽出方法：
住民基本台帳からの無作為抽出とインターネット調査モニターからの有意抽出で実施しています。
前者は調査員による訪問で調査票を配布・回収します。また、調査員による直接回収のほか、郵送又は専用回答画面（オンライン）での調査票回収を実施します。
後者はインターネット経由で調査票を配信し、専用回答画面（オンライン）で調査票を回収します。

図 3.1　環境省 HP：家庭部門の CO_2 排出実態統計調査（家庭 CO_2 統計）から抜粋 [2025/8/4 時点の掲載内容]

3.2.2　気象データ

気象データは、付録 A に示す方法により、市区町村役場の地点座標（緯度・経度）に基づき、1504 地点について 6 カ年度分（平成 29 年度～令和 4 年度）の毎時データを作成して用いた。なお、作成した地点毎の気象データには、外気温度、外気絶対湿度・相対湿度、全天日射量、大気放射量、風向・風速、降雨量等が含まれるが、本分析においてはこのうち外気温度および全天日射量の 2 項目を用いた。

3.3　分析フロー

本分析は、次の 5 つの段階から成る（図 3.2）。

- データ前処理
- 特徴量データの作成
- 分析対象世帯の選定
- 世帯間におけるエネルギー消費の差異に関する統計分析
- 機械学習を用いたエネルギー消費の差異の発生要因に関する分析

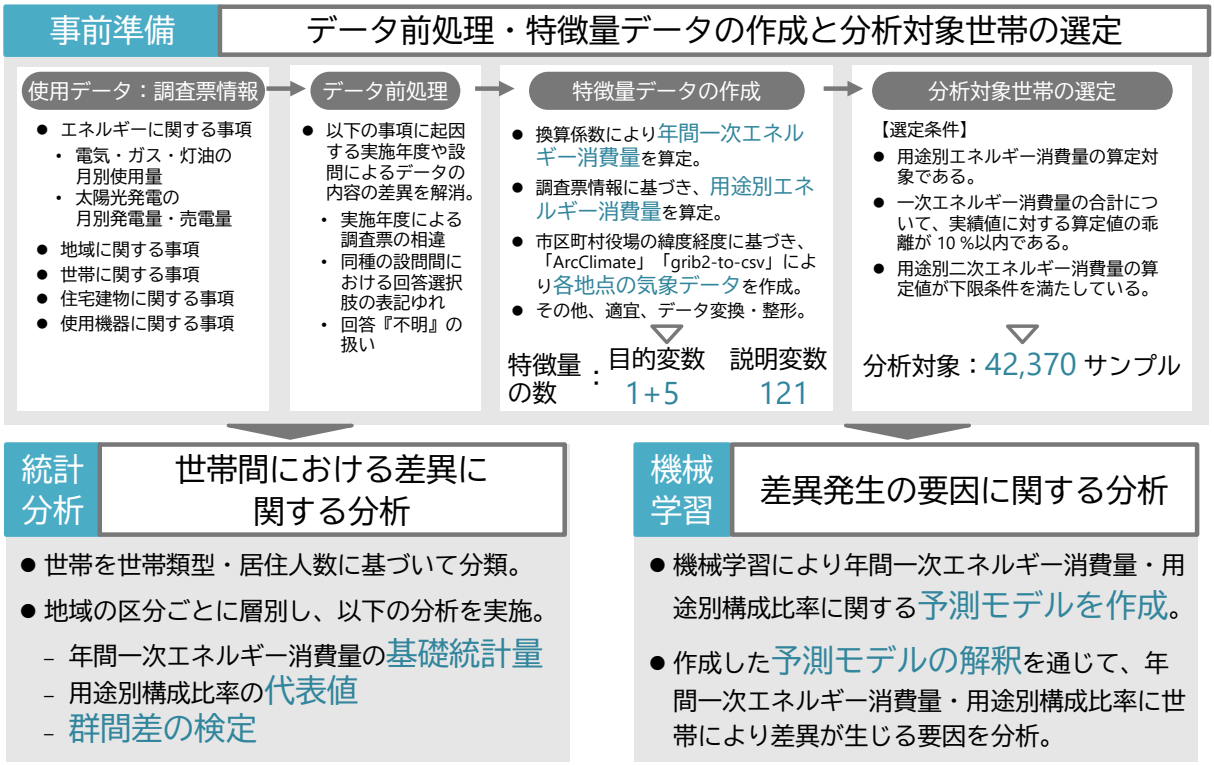


図 3.2　分析フロー

3.4　データ前処理

家庭 CO_2 統計の調査票情報には、データの内容に以下の事項に起因する実施年度や設問による差異がある。

- 実施年度による調査票の相違：

  例)「就業の有無」の回答選択肢：

  令和 2 年度調査以前：『1: 就業者』

  『2: 就業者ではない』

  令和 3 年度調査以降：『1: 就業者（在宅勤務あり）』

  『2: 就業者（在宅勤務なし）』

  『3: 就業者ではない』

- 同種の設問間における回答選択肢の表記ゆれ：

  例)「暖房の平日使用時間」と「冷房の平日使用時間」：

  「冷房の平日使用時間」には、回答選択肢に『1: 0 時間』があるが、「暖房の平日使用時間」にはこれがない。

- 『不明』として収録されている回答の意味：

  回答が『不明』として収録されている設問には、調査票に『分からない』などの対応する回答選択肢があるものと、対応する回答選択肢がないものとがある。すなわち、『不明』には“分からない”と“未回答”の二つの意味が混在している。

本分析では、上記の差異を解消してデータの扱いを簡単にするため、付録 B に示す方法により前処理を行った。

3.5　特徴量データの作成

家庭 CO_2 統計の調査票情報に対し、前掲の前処理を行った上で、付録 B に示す方法により特徴量データの作成を行った。特徴量の一覧を表 3.1 の通りに示す。なお、表中の「特徴量の名称」欄において、背景色（薄いベージュ色）を付した特徴量は、他の特徴量の作成やデータの集計・統計分析に用い、機械学習では学習させるパラメータとして使用していない特徴量である（詳細については、付録 B を参照のこと）。

表 3.1　特徴量の一覧　その 1

大分類	小分類	特徴量の名称
地域	自治体名	自治体名
	都市階級	都市階級
	気候	地域の区分、暖房度日（HDD18-18）、冷房度日（CDD24-24）、給湯度日（WDD40-40）、年積算全天日射量
世帯	世帯構成	居住人数、世帯主から見た続柄、世帯員の年代、世帯類型
	経済状況	世帯収入、就業の有無、就業割合
	住まい方	平日在宅
住宅建物	建て方	建て方
	建築時期	建築時期
	所有関係	所有関係
	住宅規模	延床面積、居室数
	窓仕様	二重サッシまたは複層ガラスの有無
暖房	使用有無（全般）	暖房機器の使用有無
	使用有無（セントラル暖房）	セントラル暖房の使用有無、セントラル暖房のエネルギー源
	使用有無（床暖房）	使用している床暖房の種類、床暖房の使用有無、電気床暖房／ガス温水床暖房／灯油温水床暖房の使用有無
	使用有無（太陽熱）	太陽熱利用暖房システムの使用有無
	使用台数（その他）	エアコン／電気ストーブ類／電気カーペット・こたつ／電気蓄熱暖房器／ガスストーブ類／灯油ストーブ類／木質系燃料ストーブ類の使用台数
	使い方	暖房の仕方、主暖房機器、主暖房機器の設定温度、主暖房機器の強弱設定、暖房の平日使用時間、暖房居室数、ペットのための暖房使用
冷房	使用有無	冷房機器の使用有無
	使用台数	冷房機器の使用台数
	種類	1／2／3／4／5 台目エアコンの種類
	製造時期	1／2／3／4／5 台目エアコンの製造時期
	使い方	冷房の平日使用時間、冷房の設定温度、ペットのための冷房使用
照明	使用有無（全般）	場所ごとに使用している照明の種類
	使用有無（居間）	ランプの使用有無、白熱電球／蛍光灯／LED／その他のランプの使用有無
	使用有無（食卓）	ランプの使用有無、白熱電球／蛍光灯／LED／その他のランプの使用有無
	使用有無（個室）	ランプの使用有無、白熱電球／蛍光灯／LED／その他のランプの使用有無
	使用有無（その他の場所）	ランプの使用有無、白熱電球／蛍光灯／LED／その他のランプの使用有無
	使い方	主照明の種類、居間照明の使用時間、調光、消灯

表 3.1　特徴量の一覧　その 2

大分類	小分類	特徴量の名称
給湯	使用有無	使用している給湯器・給湯システムの種類、給湯機の使用有無、電気ヒートポンプ式給湯機／電気温水器／ガス給湯機・ガス風呂がま／ガス小型瞬間湯沸器／灯油給湯機・灯油風呂がま／太陽熱利用給湯機／ガスエンジン発電・給湯機／家庭用燃料電池／その他給湯機の使用有無
	使い方	浴槽に湯をはる日数（夏季／冬季）、浴槽に湯をはらず、シャワーだけを使用する日数（夏季／冬季）、浴槽に湯をはらず、シャワーも使用しない日数（夏季／冬季）、夏季湯はり割合、夏季シャワー割合、夏季入浴なし割合、冬季湯はり割合、冬季シャワー割合、冬季入浴なし割合、シャワーをこまめに止める、続けて入浴、食器洗い時の工夫、コントローラーのオフ、洗面所のお湯の使用頻度、台所のお湯の使用頻度
調理	使用有無	使用している調理コンロの種類、調理コンロの使用有無、ガスコンロ／電気コンロ／その他のコンロの使用有無、
	使い方	1 日当たりの調理食数（朝・平日／朝・休日／昼・平日／昼・休日／夜・平日／夜・休日）、1 週間当たりの調理食数、電子レンジで下ごしらえ、火力調整
テレビ	使用有無	テレビの使用有無
	使用台数	テレビの使用台数
	種類	1／2／3 台目テレビの種類
	画面サイズ	1／2／3 台目テレビの画面サイズ
	製造時期	1／2／3 台目テレビの製造時期
	使い方	テレビの平日使用時間、明るさ調整、主電源オフ
冷蔵庫	使用有無	冷蔵庫の使用有無
	使用台数	冷蔵庫の使用台数
	種類	1／2 台目冷蔵庫の種類
	庫内容積	1／2 台目冷蔵庫の庫内容積
	製造時期	1／2 台目冷蔵庫の製造時期
	使い方	温度調整、つめこみ過ぎ
家電機器等	使用台数	家電製品等の使用台数
洗濯乾燥	使用台数	乾燥機能の有無、洗濯機（乾燥機能なし）／洗濯機（乾燥機能付）／衣類乾燥機（電気）／衣類乾燥機（ガス）／浴室乾燥機の使用台数
	使い方	乾燥機能使用頻度
トイレ便座	使用台数	温水洗浄便座／暖房便座（温水洗浄機能なし）の使用台数
	使い方	温水の設定温度の調整、冬季以外暖房便座機能オフ
ペット用品	使用有無	ペットのための設備・機器の種類、ペットのための設備・機器の使用有無
太陽光発電	使用有無	太陽光発電の使用有無
	容量	太陽光発電の容量
省エネ意識	使い方	省エネ行動の実施割合
エネルギー	二次エネ	1 か月当たりの電気の購入量／太陽光発電量／太陽光発電量のうちの売電分／都市ガスの購入量／LP ガスの購入量／灯油消の購入量
	一次エネ	年間一次エネルギー消費量、 暖房比率、冷房比率、給湯比率、調理比率、照明他比率

3.6 分析対象世帯の選定

以下を満たす世帯を分析の対象世帯として選定した。

- 付録 D に示す方法による用途別一次エネルギー消費量の算定対象である。
- 家庭 CO_2 統計の調査票情報から得られる各月の電気・ガス・灯油の購入量等から換算して求めた一次エネルギー消費量（付録 C に示す方法により算定）と用途別一次エネルギー消費量（付録 D に示す方法により算定）の合計の乖離が 10%以内である。
- 用途別の各月の消費電気量・ガス消費量・灯油消費量（付録 D に示す方法により算定）がゼロ以上であり、かつ、給湯に灯油を用いる場合であって給湯の年間灯油消費量（付録 D に示す方法により算定）が 10MJ 以上である。

選定の結果、算定の対象外となった件数は、表 3.2 の通りである。また、本分析の対象世帯の件数は、表 3.3 の通りである。

表 3.2 分析対象世帯の要件を満たさない件数（算定の対象外となった件数）

分析対象世帯の要件		要件を満たさない世帯［件］						
		平成 29（2017）年度	平成 30（2018）年度	平成 31（2019）年度	令 2 年（2020）年度	令 3 年（2021）年度	令 4 年（2022）年度	全体
付録 D に示す方法による用途別一次エネルギー消費量の算定対象である		1,672	1,677	1,834	2,186	2,753	1,988	12,110
算定対象の要件	家庭 CO_2 統計において用途別エネルギー消費量（二次エネルギー）の推計対象である	988	1,005	1,062	1,612	2,211	1,198	8,076
	暖房・冷房・給湯・調理用コンロのエネルギー源が全て欠損ではない	494	522	550	998	945	634	4,143
	給湯と調理用コンロのエネルギー源がそれぞれ 1 種類である	307	298	309	178	183	164	1,439
	灯油の年間購入回数が 3 回以上である	112	122	175	112	115	125	761
	中間期に分類される月の数が 1 以上である	0	0	0	0	0	208	208
家庭 CO_2 統計の調査票情報から得られる各月の電気・ガス・灯油の購入量等から換算して求めた一次エネルギー消費量と用途別一次エネルギー消費量の合計の乖離が 10%以内である		170	147	132	122	113	131	815
用途別の各月の消費電気量・ガス消費量・灯油消費量がゼロ以上であり、かつ、給湯に灯油を用いる場合であって給湯の年間灯油消費量が 10MJ 以上である		648	617	525	560	467	537	3,354
上記の条件をすべて満す		2,439	2,396	2,462	2,851	3,311	2,630	16,089

表 3.3　対象世帯の件数

選定前／選定後の別	対象世帯［件］						
	平成 29（2017）年度	平成 30（2018）年度	平成 31（2019）年度	令 2 年（2020）年度	令 3 年（2021）年度	令 4 年（2022）年度	全体
選定前※	9,505	9,996	9,660	10,015	9,804	9,479	58,459
選定後	7,066	7,600	7,198	7,164	6,493	6,849	42,370

※選定前：調査票情報に収録されているデータの件数

3.7　世帯間におけるエネルギー消費の差異に関する統計分析

3.7.1　概要

日本は南北に長く、気候条件が地域によって大きく異なるため、エネルギー消費構造にも地域差が存在する。そこで、本分析では、世帯間におけるエネルギー消費の差異を地域の区分[3)]ごとに分析することとした。また、世帯を「世帯類型」および「居住人数」に基づいて分類した。以降、上記の分類に基づく世帯の集合を“世帯群”と記す。

世帯間におけるエネルギー消費の差異は、「年間一次エネルギー消費量」および「用途構成」の2つの観点から分析した。年間一次エネルギー消費量については、「年間一次エネルギー消費量の総量 [GJ/年]」を指標として用い、世帯群ごとに基礎統計量を算出した上で、多重比較により世帯群間の差異の統計的有意性を検証した。用途構成については、「用途別構成比率 [-]」を指標として用い、群ごとに代表値を推定した上で、その結果に基づいて多重比較を行うことで世帯群間の統計的差異の有意性を検証した。なお、用途構成の指標として、用途別の年間一次エネルギー消費量 [GJ/年]を用いることも考えられた。しかし、用途別の年間一次エネルギー消費量は、電気・ガス・灯油の月別使用量等を分解する方法（トップダウン方式）によって算出される値である（参照：第2章 図2.3、付録D)。そのため、本分析では、絶対量である用途別の年間一次エネルギー消費量 [GJ/年]ではなく、相対量である用途別構成比率 [-]を指標として用いることとした。

3.7.2　記号・添え字

(1) 記号

本節で用いる記号位は、表3.4による。

表3.4　記号

記号	意味
c	定数和
D	組成ベクトルの成分の数
$d_A(\mathbf{x},\ \mathbf{y})$	組成ベクトル $\mathbf{x}$ と組成ベクトル $\mathbf{y}$ との間のAitchison距離
F_{obs}	観測データ（元の群ラベル）から計算された pseudo-F 統計量
F_{perm}	群ラベルを置換して得られた pseudo-F 統計量
$ilr(\ \ )$	ILR変換を表す関数
$ilr^{-1}(\ \ )$	逆ILR変換を表す関数
$\mathbf{m}$	組成データの代表値
$\mathbf{m}^*$	代表値の候補
N_{perm}	置換回数
n	サンプル数
p	PERMANOVA により算出される p 値
R	分子側に割り当てられた添え字の集合

記号	意味
r	集合 R に含まれる成分の数
$\mathbb{R}^D$	D 次元実ユークリッド空間
S	分母側に割り当てられた添え字の集合
s	集合 S に含まれる成分の数
$\mathcal{S}^D$	D 次元の組成データの定義域
$\mathbf{x}$	組成ベクトル
x	組成ベクトル $\mathbf{x}$ の成分
x^*	置換後の組成ベクトルの成分
$\mathbf{y}$	組成ベクトル
y	組成ベクトル $\mathbf{y}$ の成分
$\mathbf{z}$	組成ベクトル
z	組成ベクトル $\mathbf{z}$ の成分
δ	十分に小さい正の数

(2) 添え字

本節で用いる添え字は、表 3.5 による。

表 3.5　添え字

添え字	意味
i	用途（1：暖房、2：冷房、3：給湯、4：調理、5：照明他）
k	サンプル

3.7.3　年間一次エネルギー消費量

分析は、次の手順により行った。

(1) 基礎統計量の算出

(2) 群間差の検定

具体的な方法を以下に示す。

(1) 基礎統計量の算出

地域の区分ごとに層別した上で、各世帯群について平均値、平均値の 95 %信頼区間、標準偏差、25%値、50%値、75%値、最小値、最大値を算出した。

(2) 群間差の検定

年間一次エネルギー消費量に関する群間差の検定は、地域の区分ごとに層別した上で、Steel-Dwass 検定を用いて実施した。Steel–Dwass 検定は、3 群以上を対象として、群間の分布の位置

（中央値）の差を検定するノンパラメトリックな多重比較手法である。

3.7.4　用途構成

用途構成の差異の分析は、用途別構成比率を指標として行った。ここで、用途別構成比率とは「暖房比率」「冷房比率」「給湯比率」「調理比率」「照明他比率」であり、単位を無次元とすると、各変数がいずれも 0.0～1.0 の間の正の値で、その総和は定数（1.0）となるという制約がある。このような特徴を有するデータは、「組成データ（CoDa： Compositional Data）」と総称される[4)]。

組成データは、実数の比例尺度データと同じように平均値や信頼区間などの統計処理や多変量解析などの統計解析を適用できないことが知られている。この問題は、地質学分野において古くから認識されており、組成データを適切に解析するための方法について研究[例えば4)5)]が進められた。その方法は、数学的なテクニックを基本としていることから、地質学分野以外の分析にも応用することができる。

本分析では、用途別構成比率のような組成データが持つ定数和制約および変数間の相対関係を適切に扱うため、組成データ解析（CoDA： Compositional Data Analysis）の一つである「対数比変換」を採用した[4) 5)]。対数比変換とは、ある共通の変数でその他の変数を規格化し、さらにこれを対数変換することにより、組成データを単体空間（組成データの標本空間）から実空間に写像する方法である。これにより、実数に適用される演算や統計学的手法などを適用して解析を適用することが可能となる。対数変換にはいくつか種類があるが、本分析では、統計分析においては、ILR（Isometric Log-Ratio）変換（等長対数比変換）[5)] を採用した。

分析は、次の手順により行った。

（1）ゼロ置換（対数比変換の前処理）

（2）ILR（Isometric Log-Ratio）変換

（3）用途別構成比率の代表値の推定

（4）群間差の検定

具体的な方法を以下に示す。

（1）ゼロ置換（対数比変換の前処理）

対数比変換の前処理として、ゼロ値の置換を行った。ゼロ値は、丸め誤差に起因するゼロ値（rounded zero）としてみなし、Martín-Fernández らによる乗法置換法（multiplicative replacement）[6)] を適用した。具体的には、組成ベクトル $\mathbf{x}$ の成分 x_i は、ゼロ成分か否かに応じて式（1）に従い置換を行った。なお、本分析では成分間のバランスを考慮して、置換後の値は成分ごとに異なる値を設定した。

$$x_i^* = \begin{cases} \delta_i & if\ x_i = 0 \\ \left(1 - \dfrac{\sum_{k|x_k=0} \delta_k}{c}\right) x_i & if\ x_i > 0 \end{cases} \tag{1a}$$

$$\delta_i = \begin{cases} 10^{-5} & (i = 1) \\ 10^{-6} & (i = 2) \\ 10^{-5} & (i = 3) \\ 10^{-6} & (i = 4) \\ 10^{-4} & (i = 5) \end{cases} \tag{1b}$$

ここで、

c　：定数和（= 1.0）

x_i　：組成ベクトル $\mathbf{x}$ の 成分

x_i^*　：置換後の組成ベクトルの成分

δ　：十分に小さい正の数 [-]

添え字 i　：用途（1：暖房、2：冷房、3：給湯、4：調理、5：照明他）

である。

（2）ILR（Isometric Log-Ratio）変換

組成データを ILR 変換（等長対数比変換）により単体空間から実空間に写像した。ILR 変換は、成分間の対数比に基づく等長変換（変換前後において、どの 2 点間の距離も変えない変換）であり、D 次元の組成ベクトルを $D-1$ 次元の実数ベクトルとして表現する。

Egozcue ら[5] による正規直交基底に基づく ILR 変換の第 i 成分 $ilr_i(\boldsymbol{x})$ は、逐次二分構造に従い、式（2）のように表される。

$$ilr_i(\boldsymbol{x}) = \sqrt{\frac{r_i s_i}{r_i + s_i}} log \frac{\left(\prod_{j \in R_i} x_j\right)^{1/r_i}}{\left(\prod_{j \in S_i} x_j\right)^{1/s_i}} \qquad (i = 1, \ \ldots, \ D-1) \tag{2a}$$

$$r_i = |R_i| \tag{2b}$$

$$s_i = |S_i| \tag{2c}$$

$$R_i = \{i\} \tag{2d}$$

$$S_i = \{i+1, \ \ldots, \ D\} \tag{2e}$$

ここで、

D　：組成ベクトル $\boldsymbol{x}$ の 成分数

$ilr(\ \ )$　：ILR変換を表す関数

R　：分子側に割り当てられた添え字の集合

r　　：集合 R_i に含まれる成分の数

S　　：分母側に割り当てられた添え字の集合

s　　：集合 S_i に含まれる成分の数

x_i　　：組成ベクトル X の 成分

添え字 i ：用途（1：暖房、2：冷房、3：給湯、4：調理、5：照明他）

である。

本分析のように 5 成分からなる組成ベクトル $\boldsymbol{x} = (x_1,\ x_2,\ x_3,\ x_4,\ x_5)$ （ただし、$x_i > 0$ ）を例にすると、ILR 変換の第 i 成分 $ilr_{\mathrm{i}}(\boldsymbol{x})$ は、式（3）のように表される。

$$ilr_1(\boldsymbol{x}) = \sqrt{\frac{4}{5}}\left(log\, x_1 - \frac{1}{4}\sum_{j=2}^{5} log\, x_j \right) \tag{3a}$$

$$ilr_2(\boldsymbol{x}) = \sqrt{\frac{3}{4}}\left(log\, x_2 - \frac{1}{3}\sum_{j=3}^{5} log\, x_j \right) \tag{3b}$$

$$ilr_3(\boldsymbol{x}) = \sqrt{\frac{2}{3}}\left(log\, x_3 - \frac{1}{2}\sum_{j=4}^{5} log\, x_j \right) \tag{3c}$$

$$ilr_4(\boldsymbol{x}) = \sqrt{\frac{1}{2}}(log\, x_4 - log\, x_5) \tag{3d}$$

ここで、

$ilr(\ \ )$　：ILR変換を表す関数

x_i　　：組成ベクトル $\boldsymbol{x}$ の 成分

添え字 i ：用途（1：暖房、2：冷房、3：給湯、4：調理、5：照明他）

である。この変換により、組成データに特有の定数和制約を解消し、標準的な多変量解析手法の適用が可能となる。

（3）用途別構成比率の代表値の推定

組成データの代表値には、Pawloesky-Glahn and Egozcue[7) 8)] による，Aitchison 幾何における中心を用いる。Aitchison 幾何における中心は、成分間の比構造を保持したまま、Aitchison 距離の二乗和が最小となる中心点として定義されものであり、実空間における平均値に相当する概念である。

Aitchison 距離 $d_A(\boldsymbol{x},\ \mathbf{y})$ は、組成データ間の距離を測るために定義された距離尺度であり、D 成分からなる組成ベクトル $\boldsymbol{x}$、 $\mathbf{y}$ に対して、対数比に基づく表現として式（4）のように表される。

$$d_A(\boldsymbol{x},\ \boldsymbol{y}) = \left[\frac{1}{2D}\sum_{i=1}^{D}\sum_{j=1}^{D}\left\{log\left(\frac{x_i}{x_j}\right) - log\left(\frac{y_i}{y_j}\right)\right\}^2\right]^{\frac{1}{2}} \tag{4a}$$

$$\boldsymbol{x} = (x_1,\ \cdots,\ x_D) \tag{4b}$$

$$\boldsymbol{y} = (y_1,\ \cdots,\ y_D) \tag{4c}$$

ここで、

D　　：組成ベクトル $\boldsymbol{x}$ および 組成ベクトル $\boldsymbol{y}$の成分の数

$d_A(\boldsymbol{x},\ \boldsymbol{y})$　　：組成ベクトル $\boldsymbol{x}$ と組成ベクトル $\boldsymbol{y}$ との間の Aitchison 距離

$x_i,\ y_i$　　：組成ベクトル $\boldsymbol{x}$ および 組成ベクトル $\boldsymbol{y}$の成分

$\boldsymbol{x},\ \boldsymbol{y}$　　：組成ベクトル

添え字 i　　：用途（1：暖房、2：冷房、3：給湯、4：調理、5：照明他）

である。

このとき、ILR 変換により組成ベクトルをユークリッド空間へ写像すると、Aitchison 距離 $d_A(\boldsymbol{x},\ \boldsymbol{y})$は等長写像によりユークリッド距離と一致し、式（5）のように表される。

$$d_A(\boldsymbol{x},\ \boldsymbol{y}) = \|ilr(\boldsymbol{x}) - ilr(\boldsymbol{y})\|_2 \tag{5}$$

ここで、

$d_A(\boldsymbol{x},\ \boldsymbol{y})$　　：組成ベクトル $\boldsymbol{x}$ と組成ベクトル $\boldsymbol{y}$ との間の Aitchison 距離

$ilr(\ )$　　：ILR 変換を表す関数

$\boldsymbol{x},\ \boldsymbol{y}$　　：組成ベクトル

である。$\|\ \ \|_2$はユークリッドノルム（Euclidean norm, L2 ノルム）という数学表現であり、一般化すると式（6）のように表される。

$$\|\boldsymbol{z}\|_2 = \sqrt{\sum_{k=1}^{K} z_k^2}\quad (\boldsymbol{z} \in \mathbb{R}^K) \tag{6}$$

ここで、

$\boldsymbol{z}$　　：K 次元ベクトル

z_k　　：K 次元ベクトル $\boldsymbol{z}$ の成分

$\mathbb{R}^K$　　：K 次元実ユークリッド空間

である。

ILR 変換により組成データはユークリッド空間へ等長に写像され、Aitchison 距離はユークリッド距離として表される。この性質により、組成データに対する代表値は、ユークリッド空間における二乗距離の総和を最小化する点として定義することができる。

従って、組成データの代表値 **m** は、組成ベクトル **x** と代表値の候補 $\mathbf{m}^*$ との Aitchison 距離

の二乗和が最小となる点として定義され、式（7）のように表される。

$$\boldsymbol{m} = arg \min_{\boldsymbol{m}^* \in \mathcal{S}^D} \|ilr(\boldsymbol{x}_k) - ilr(\boldsymbol{m}^*)\|_2^2 \tag{7}$$

ここで、

n　　：サンプル数

$ilr(\ \ )$　：ILR 変換

$\mathbf{m}$　　：組成データの代表値

$\mathbf{m}^*$　　：代表値の候補

$\mathbf{x}_k$　　：サンプル k の組成ベクトル

$\mathcal{S}^D$　　：D 次元の組成データの定義域（正の成分を持ち、成分和が 1 に正規化されたD 次元組成ベクトル全体からなる単体空間）

である。このとき、ユークリッド距離の二乗和を最小化する問題の解は算術平均に一致するため、中心の ILR 座標は算術平均として与えられる。

算術平均は、一般化すると式（8）のように表される。

$$\bar{\boldsymbol{z}} = \frac{1}{n}\sum_{k=1}^{n} \boldsymbol{z}_k \quad (\boldsymbol{z}_k \in \mathbb{R}^K) \tag{8}$$

ここで、

$\mathbf{z}$　　：K 次元ベクトル

$\bar{\mathbf{z}}$　　：K 次元ベクトル $\mathbf{z}$ の平均

$\mathbb{R}^K$　　：K 次元実ユークリッド空間

である。従って、式（7）は式（9）のように表される。

$$ilr(\boldsymbol{m}) = \frac{1}{n}\sum_{k=1}^{n} ilr(\boldsymbol{x}_k) \tag{9}$$

ここで、

n　　：サンプル数

$ilr(\ \ )$　：ILR 変換を表す関数

$\boldsymbol{m}$　　：組成データの代表値

$\mathbf{x}_k$　　：サンプル k の組成ベクトル

である。組成データの代表値 $\boldsymbol{m}$ は、式（9）を逆 ILR 変換すること得られ、式（10）のように表される。

$$\boldsymbol{m} = ilr^{-1}\left(\frac{1}{n}\sum_{k=1}^{n} ilr(\boldsymbol{x}_k)\right) \tag{10}$$

ここで、

n　　：サンプル数

$ilr(\ \ )$　：ILR 変換を表す関数

$ilr^{-1}(\ \ )$：逆 ILR 変換を表す関数

$\mathbf{m}$　　：組成データの代表値

$\mathbf{x}_k$　　：サンプル k の組成ベクトル

である。

ただし、組成データの代表値は、通常の平均と同様に、外れ値や異常値の影響を受けやすい。特に多次元の組成空間においては、少数の外れた観測が推定された代表組成を大きく歪める恐れがある。そこで本分析では、代表値の推定にロバスト推定を導入し、外れ値の影響を抑制した。

具体的には、最小共分散行列式（Minimum Covariance Determinant; MCD）法による推定を基本とした。MCD 法は、共分散行列式が最小となる観測値の部分集合（サンプル数 $\alpha \times n$）を用いて中心位置と散布を特定する手法であり、多変量空間における外れ値の検出と抑制に優れている。

しかし、MCD 法はデータの次元数に対してサンプル数が不十分な場合、計算が不安定または不能になるという制約がある。そのため、サンプル数が不足する場合には代替手法としてトリム平均を用いた。トリム平均は、各成分の分布の両端から一定割合（トリミング率）を除外して平均を算出する手法である。多変量的な相関は考慮しないものの、計算が容易で解釈性に優れ、小標本においても安定したロバスト推定が可能となる。

具体的には、表 3.6 に示すように、サンプル数 n と次元数 $D-1$ に応じて推定手法を選択した。ここで、MCD 法における観測値の部分集合の割合 α については、複数の水準（0.70、0.75、0.80）を用いた感度分析を実施した。その結果、0.80 から 0.75 への変更では推定結果に大きな変化が生じた一方、0.75 から 0.70 への変更による差は比較的小さく、推定値が安定化することが確認された。このため、本分析では推定結果の安定域を代表する値として、$n \geq 2 \times (D-1)$ の場合には $\alpha = 0.75$ を、$D-1 < n < 2 \times (D-1)$ の場合にはより保守的な $\alpha = 0.70$ を採用した。なお、表 3.6 に示す考え方に基づいて推定を行った結果、8 つの世帯群についてはトリム平均が、それ以外については MCD 法が適用された。

表 3.6　組成データの代表値の算出における
ロバスト推定手法の選択方法および選択結果

サンプル数nと 次元数$D-1$の関係	推定手法の概要	選択結果
$n > 20$ かつ $n \geq 2 \times (D-1)$	• 次元数に対して十分なサンプル数が確保されていると判断し、MCD 法を適用。 • α は、感度分析の結果に基づいて推定結果の安定域を代表する値（0.75）に設定。	選択されず
$n > 20$ かつ $D-1 < n < 2 \times (D-1)$	• MCD 法を適用。 • 推定の安定性が低下する可能性を考慮し、α は、より保守的な値（0.70）に設定。	下段の 8 群を除く 全ての世帯群
$n \leq 20$ または $n < 2 \times (D-1)$	• MCD 法の適用が難しいため、トリム平均を採用。 • トリミング率は、0.10 に設定。	1 地域_単独(非高齢)、 1 地域_単独(高齢)、 1 地域_夫婦と子(未成年)4 人 1 地域_夫婦と子(未成年)5 人以上、 1 地域_夫婦と子(成年)4 人、 1 地域_不明、 3 地域_夫婦と子(未成年)5 人以上、 4 地域_夫婦と子(未成年)5 人以上

各代表値の 95% 信頼区間は、ブートストラップ法により求めた。具体的には、元データからサンプル数 n と同じ大きさの再標本を置換抽出し、500 回反復して各再標本の 組成データの代表値を計算した。この分布の 2.5 パーセンタイルおよび 97.5 パーセンタイルを下限・上限として 95% 信頼区間とした。

（4）群間差の検定

用途別構成比率に関する群間差の検定は、地域の区分ごとに層別した上で、ILR 変換後のデータに対して PERMANOVA（permutational multivariate analysis of variance：置換多変量分散分析）を適用することにより実施した。PERMANOVA は距離行列に基づいて群間差を検定するノンパラメトリック手法である。PERMANOVA により算出される p 値は、観測された pseudo-F 統計量以上の値が群ラベルをランダムに置換して得られた pseudo-F 分布において出現する割合であり、式（11）で表される。

$$p = \frac{1 + \left|\{F_{perm} | F_{perm} > F_{obs}\}\right|}{1 + N_{perm}} \tag{11}$$

ここで、

p　　：PERMANOVA により算出される p 値

F_{obs}　　：観測データ（元の群ラベル）から計算された pseudo-F 統計量

F_{perm}　　：群ラベルを置換して得られた pseudo-F 統計量

N_{perm}　　：置換回数

である。従って、PERMANOVA により算出される p 値の最小値は、$1/(1+N_{perm})$ である。

本分析では、距離として、Aitchison 距離を用いた。置換回数は、9,999 回とした。また、複数の群間比較に伴う第 I 種過誤の増大を抑制するため、p 値の補正には Holm 法を用いた。

3.8　機械学習を用いたエネルギー消費の差異の発生要因に関する分析

3.8.1　概要

本分析で用いた家庭 CO_2 統計のような調査票に基づく政府統計データでは、未回答に起因する欠損が避けられない。このようなデータに対して統計解析を適用する場合、完全ケース分析（欠損を含むサンプルを除外し、欠損のないサンプルのみを用いて分析を行う手法）や欠損補完といった前処理が不可避となり、解析対象の減少や補完方法に依存した不確実性が生じうる。一方、機械学習手法は、欠損値を許容するアルゴリズムや、欠損の有無や発生構造そのものを特徴量として取り込む枠組みを有しており、欠損を含む実データの特性を保持したまま、要因構造を探索的に把握できるという利点がある。

そこで、本分析では、機械学習を適用し、家庭 CO_2 統計の調査票情報から得られる住宅、世帯、保有機器とその使い方などを説明変数として、年間一次エネルギー消費量および用途別構成比率を予測するモデルを作成し、そのモデルの解釈を通じて世帯間にエネルギー消費の差異が生じる要因を分析した。

3.8.2　記号・添え字

（1）記号

本節で用いる記号は、表 3.7 による。

表 3.7　記号および単位

記号	意味
	---------- SHAP ----------
$f(\ \ )$	予測値
F_g	要因 g に属する特徴量の集合
H_{h}	分析単位 h に属するサンプルの集合
I	特徴量重要度
M	特徴量（説明変数）の総数
n	サンプル数
$\mathbf{x}$	特徴量ベクトル
ϕ	寄与度（SHAP 値）
ϕ_0	予測値の期待値
$\bar{\phi}$	平均 SHAP 値
$\overline{\phi^*}$	基準化平均 SHAP 値
	---------- - CLR 変換 ----------
$clr(\ \ )$	CLR変換を表す関数
D	組成ベクトルの成分の数
$\mathbf{x}$	組成ベクトル
x	組成ベクトル $\mathbf{x}$ の成分

（2）添え字

本節で用いる添え字は、表 3.8 による。

表 3.8　添え字

添え字	意味
	---------- SHAP ----------
g	要因
h	分析単位
j	特徴量
k	サンプル
	---------- - CLR 変換 ----------
i	用途（1：暖房、2：冷房、3：給湯、4：調理、5：照明他）

3.8.3　年間一次エネルギー消費量

分析は、次の手順により行った。

（1）機械学習モデルの作成

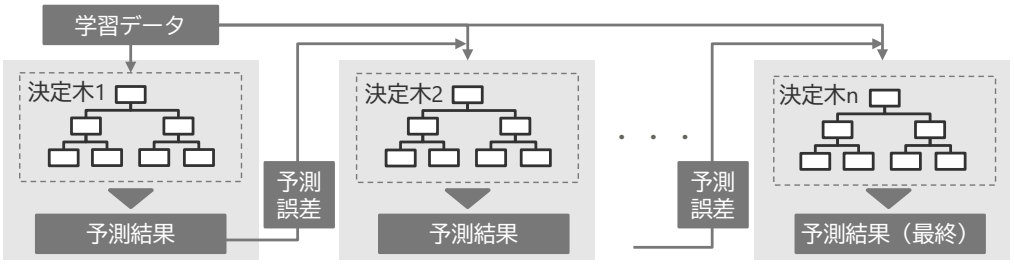
（2）予測モデルによる予測値の算出

（3）予測モデルの解釈および要因分析

具体的な方法を以下に示す。

（1）機械学習による予測モデルの作成

機械学習モデルとして勾配ブースティング木（Gradient Boosting Decision Tree、GBDT）（図 3.3）を用い、家庭 CO_2 統計の調査票情報から得られる住宅、世帯、保有機器とその使い方などを説明変数として年間一次エネルギー消費量を予測するモデルを作成した。



GBDT は、次の 3 つの手法が組み合わされた機械学習の手法。

- 勾配降下法（Gradient Descent）：前のモデルで間違ったデータ点や予測誤差の大きかったデータ点を重点的に学習。
- アンサンブル学習（Boosting）：弱学習器（ここでは、決定木）を並列に多数組み合わせることで、強力な予測モデルを構築。
- 決定木（Decision Tree）：特徴量の非線形な関係や複雑な相互作用を捉えることが可能。

図 3.3　勾配ブースティング木（Gradient Boosting Decision Tree、GBDT）の概念図

GBDT のアルゴリズムには、勾配ブースティング決定木の代表的なアルゴリズムである、XGBoost（eXtreme Gradient Boosting）を用いた。XGBoost の主なハイパーパラメータを表 3.9 に示す。

表 3.9　XGBoost の主なハイパーパラメータ

カテゴリ	役割	変数名	変数名（日本語）
モデル構造・学習方式	モデルの基本構造 および木構築方法を規定する	booster	ブースティング方式
		tree_method	木構築方式
		grow_policy	木成長方針
		sampling_method	サンプリング方式
モデル容量・学習制御	モデルの表現力 および学習の進み方を制御する	n_estimators	木の本数
		learning_rate	学習率
		max_depth	木の最大深さ
		min_child_weight	最小子ノード重み
サンプリング設定	過学習抑制のための データ・特徴量サンプリング	subsample	行サンプリング率
		colsample_bytree	列サンプリング率
正則化	モデルの複雑さを抑え、 汎化性能を高める	gamma	分割最小損失減少量
		reg_alpha	L1 正則化係数
		reg_lambda	L2 正則化係数
不均衡対応・初期値	学習初期条件や データ不均衡に対応	scale_pos_weight	クラス重み調整係数
		base_score	初期予測値
目的関数・評価指標	最適化対象および 評価基準を規定	objective	目的関数
		eval_metric	評価指標
再現性の確保	乱数制御による 再現性担保	random_state	乱数シード
		seed	学習用シード値
		seed_per_iteration	反復ごとのシード設定

XGBoost のハイパーパラメータの調整には、データセット全体（全分析対象世帯のデータ）を使用し、5 分割交差検証（5-fold cross-validation）によるモデル評価を基に、ランダムサーチによる探索を行った（図 3.4）。探索回数は 500 回とし、最も良好な性能を示したハイパーパラメータの組合せを最終的に採用した。モデルの残差評価および性能指標には RMSE（root mean squared error）を用いた。なお、説明変数に含まれる欠損値については、事前の補完処理を行わず、XGBoost の内部アルゴリズムによるに内部処理に委ねた。

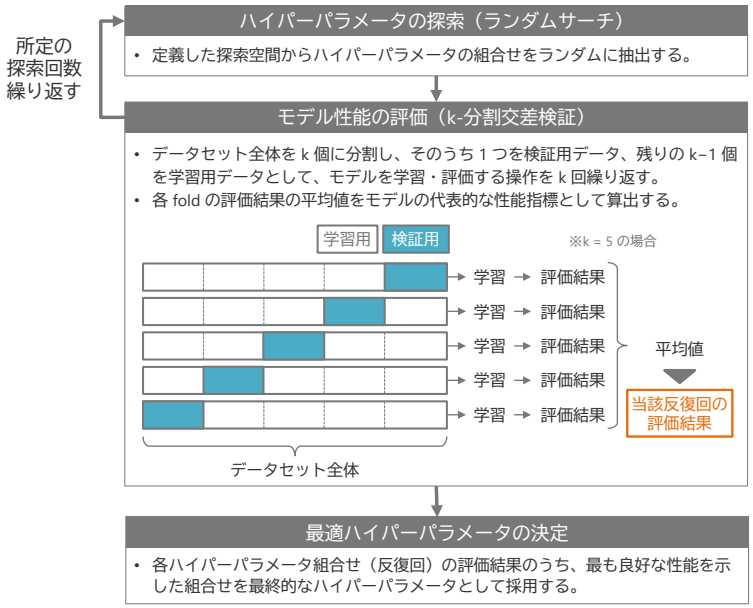


図 3.4　ハイパーパラメータの調整フロー

5 分割交差検証によるモデルの評価結果およびハイパーパラメータの調整結果をそれぞれ表 3.10 および表 3.11 に示す。決定したハイパーパラメータを用いて、全分析対象世帯を対象に再学習を行った。再学習により得られたモデルを最終的な予測モデルとした。

表 3.10　5 分割交差検証によるモデルの評価結果

指標	単位	平均	標準偏差
MAE (Mean Absolute Error, 平均絶対誤差)	GJ/年	13.24	0.102
MSE (Mean Squared Error, 平均二乗誤差)	GJ^2/$年^2$	341.13	9.923
RMSE (Root Mean Squared Error, 平均二乗誤差の平方根)	GJ/年	18.47	0.269
R^2 (決定係数)	-	0.72	0.011
RMSLE (Root Mean Squared Logarithmic Error, 平均二乗対数誤差の平方根)	-	0.25	0.003
MAPE (Mean Absolute Percentage Error, 平均絶対パーセント誤差)	-	0.22	0.002

表 3.11　ハイパーパラメータの調整結果

変数名	変数名（日本語）	値（調整後）
booster	ブースティング方式	gbtree
tree_method	木構築方式	auto
grow_policy	木成長方針	depthwise
sampling_method	サンプリング方式	uniform
n_estimators	木の本数	270
learning_rate	学習率	0.05
max_depth	木の最大深さ	7
min_child_weight	最小子ノード重み	1
subsample	行サンプリング率	0.9
colsample_bytree	列サンプリング率	0.5
gamma	分割最小損失減少量	0
reg_alpha	L1 正則化係数	0.15
reg_lambda	L2 正則化係数	0.005
scale_pos_weight	クラス重み調整係数	44.2
base_score	初期予測値	68.24
objective	目的関数	reg:squarederror
eval_metric	評価指標	rmse
random_state	乱数シード	1
seed	学習用シード値	1
seed_per_iteration	反復ごとのシード設定	0

（2）予測モデルによる予測値の算出

（1）で作成した予測モデルを用いて、年間一次エネルギー消費量の予測値を算出した。参考として、世帯群ごとに予測値の統計量を算出した結果を補足 II に掲載する。

本分析では、ここで算出した予測値をモデル解釈および要因分析の対象とした。

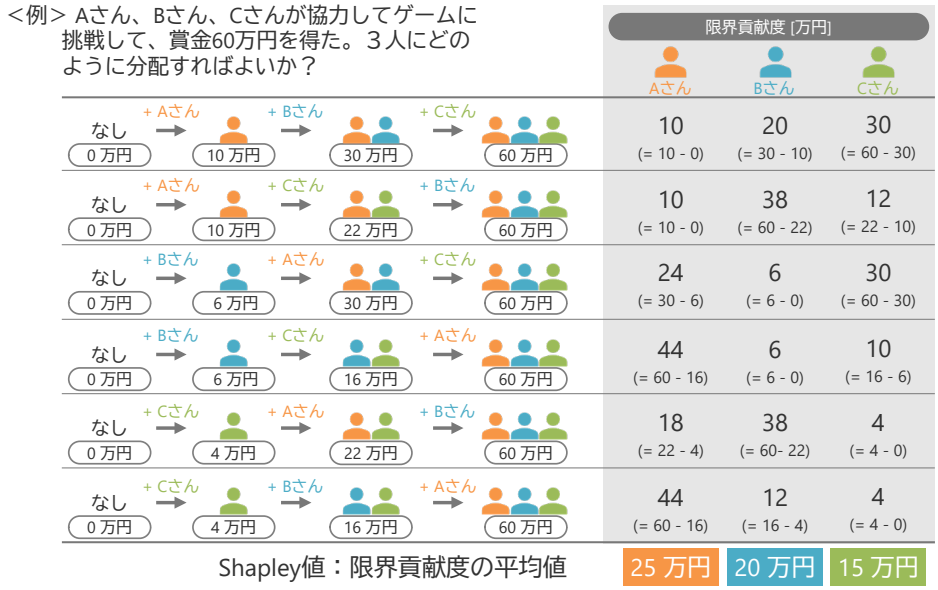
（3）予測モデルの解釈および要因分析

（2）で算出した予測結果を SHAP（SHapley Additive exPlanations）[9] により解釈し、その結果に基づいて次の分析を行った。

- 特徴量ごとに SHAP 値の絶対値の平均（特徴量重要度）を算出することで、予測に対する各特徴量の平均的な寄与度を分析。
- 特徴量ごとの SHAP 値の分布を可視化することにより、各特徴量が年間一次エネルギー消費量の予測に与える影響の大きさや方向性を分析。
- 特徴量をその内容的な性質に基づいて複数の要因に分類した上で、各要因に属する特徴量の SHAP 値の平均値を算出し、要因ごとの予測における寄与の方向性と大きさを分析。

ここで、本分析で用いた SHAP は、協力ゲーム理論の Shapley 値の概念を応用し、機械学習モデルによる予測における各特徴量（説明変数）の寄与度を算出する手法である。Shapley 値は、協力ゲームにおいて得られる報酬を各プレイヤーの貢献度に応じて公平に分配するための指標の一つである（図 3.5）。機械学習モデルにおいては、「報酬の総額」が「予測値と期待値の差分」に、「プレイヤー」が「特徴量（説明変数）」にそれぞれ対応づけられる。ただし、SHAP は「機械学習モデルがなぜそのような予測を出したのか」を解釈する手法であり、因果関係を直接的に示すものではない。

＜例＞ A さん、B さん、C さんが協力してゲームに挑戦して、賞金 60 万円を得た。
３人にどのように分配すればよいか？



DataRobot 社：「SHAP を用いて機械学習モデルを説明する」を参考に作成。
https://www.datarobot.com/jp/blog/explain-machine-learning-models-using-shap/

図 3.5　Shapley 値の概念図

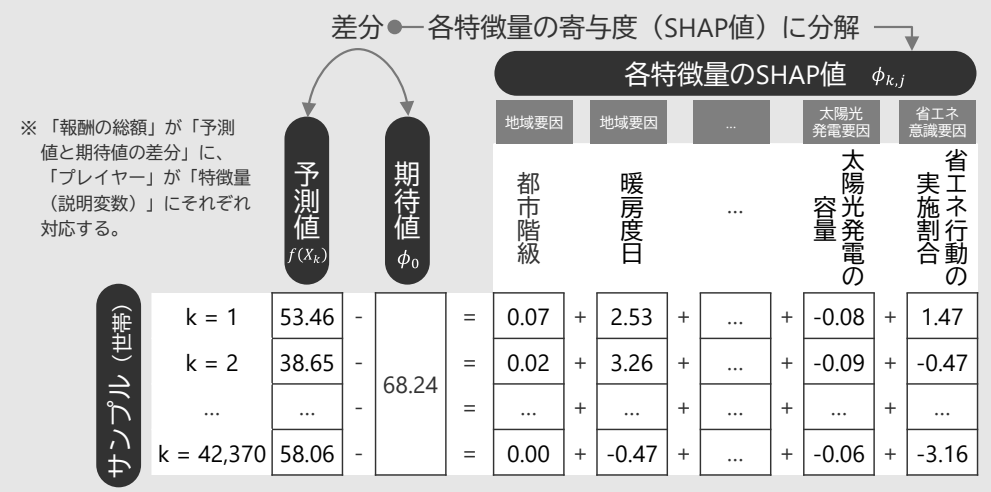


図 3.6　SHAP による特徴量の寄与度（SHAP 値）算出の概念図

SHAP は、サンプル k において予測値 $f(X_k)$ と予測値の期待値 ϕ_0 との間に生じた差分を各特徴量に対してその寄与度に応じて配分する（図 3.6）。すなわち、SHAP では予測値 $f(\mathbf{x}_k)$ は式（12）のように分解される。

$$f(\boldsymbol{x}_k) = \phi_0 + \sum_{j}^{M} \phi_{k,j} \qquad (12)$$

ここで、

M　　：特徴量（説明変数）の総数

$f(\boldsymbol{x}_k)$　　：サンプル k の予測値

$\phi_{k,j}$　　：サンプル k の特徴量 j の寄与度（SHAP 値）

ϕ_0　　：予測値の期待値

$\boldsymbol{x}_k$　　：サンプル k の特徴量ベクトル

である。

SHAP 値には加法性があり、各特徴量の SHAP 値を合計すると、期待値からの予測値の差分に一致する。この性質により、SHAP 値をサンプル単位で評価するだけでなく、特定の世帯類型などの属性群に属するサンプルに限定して集計することで、群間における寄与構造の差異を分析することが可能である。

特に、特徴量ごとに SHAP 値の絶対値を平均した指標は特徴量重要度と呼ばれ、予測における各特徴量の平均的な寄与の大きさを表す指標として広く用いられている。特徴量 j の特徴量重要度 I_j は、式（13）により表される。

$$I_j = \frac{1}{n}\sum_{k}^{n}|\phi_{k,j}| \tag{13}$$

ここで、

n　　：サンプル数

I_j　　：特徴量 j の特徴量重要度

$\phi_{k,j}$　　：サンプル k の特徴量 j の寄与度（SHAP 値）

である。

また、本分析では、世帯群ごとに特徴量の SHAP 値の平均値を算出し、各世帯群における特徴量ごとの予測における寄与の方向性と大きさを評価した。分析単位（本分析では、世帯類型もしくは世帯群） h の特徴量 j の平均 SHAP 値$\bar{\phi}_{h,j}$は、式（14）により表される。

$$\bar{\phi}_{h,j} = \frac{1}{|H_h|}\sum_{k\in H_h}\phi_{k,j} \tag{14}$$

ここで、

n　　：サンプル数

$\phi_{k,j}$　　：サンプル k の特徴量 j の SHAP 値

$\bar{\phi}_{h,j}$　　：分析単位 h の特徴量 j の平均 SHAP 値

H_{h}　　：分析単位 h に属するサンプルの集合

である。

さらに、特徴量をその内容的な性質に基づいて複数の要因に分類した上で、各要因に属する特徴量の SHAP 値の平均値を算出し、要因ごとの予測における寄与の方向性と大きさを評価した。

分析単位（本分析では、世帯類型もしくは世帯群） h の要因 g の平均 SHAP 値$\bar{\phi}_{h,g}$は、式（15）により表される。

$$\bar{\phi}_{h,g} = \frac{1}{|H_h|} \sum_{k \in H_h} \bar{\phi}_{k,g} \tag{15a}$$

$$\bar{\phi}_{k,g} = \frac{1}{|F_g|} \sum_{j \in F_g} \phi_{k,j} \tag{15b}$$

ここで、

n　　：サンプル数

I_j　　：特徴量 j の特徴量重要度

$\phi_{k,j}$　　：サンプル k の特徴量 j の SHAP 値

$\bar{\phi}_{k,g}$　　：サンプル k の要因 g の平均 SHAP 値

$\bar{\phi}_{h,g}$　　：分析単位 h の要因 g の平均 SHAP 値

F_g　　：要因 g に属する特徴量の集合

H_{h}　　：分析単位 h に属するサンプルの集合

である。

分析単位 h の要因 g の平均 SHAP 値$\bar{\phi}_{k,f}$は、方向性（正負の符号）を保持したまま寄与の大きさを示す指標であるが、その絶対値の総和は分析単位ごとに異なる。このため、各分析単位において、要因ごとの平均 SHAP 値の絶対値の総和を算出し、それぞれの平均 SHAP 値をその総和で除することにより基準化を行った。分析単位 h の要因 g の基準化平均 SHAP 値$\bar{\phi}^*_{h,g}$は、式（16）により表される。

$$\bar{\phi}^*_{h,g} = \frac{\bar{\phi}_{h,g}}{\sum_a |\bar{\phi}_{h,a}|} \tag{16}$$

ここで、

$\bar{\phi}_{h,g}$　　：分析単位 h の要因 g の平均 SHAP 値

$\bar{\phi}^*_{h,g}$　　：分析単位 h の要因 g の基準化平均 SHAP 値

である。上記の操作を行うことにより、式（17）が成立することになる。

$$\sum_g |\bar{\phi}^*_{h,g}| = 1 \tag{17}$$

ここで、

$\bar{\phi}^*_{h,g}$　　：分析単位 h の要因 g の基準化平均 SHAP 値

である。すなわち、基準化平均 SHAP 値は、各要因の寄与方向（正負）を保持しつつ、要因間の相対的な寄与割合を明示しており、寄与構造を分析単位間で比較することが可能となる。

3.8.4　用途構成

用途構成の差異の発生要因に関する分析は、用途別構成比率を目的変数とした機械学習モデルを作成し、そのモデルを解釈することにより行った。用途別構成比率に対しては、前項の統計分析と同様に対数比変換を行った。ただし、機械学習においては、CLR（Centered Log-Ratio）変換（中心対数比変換）[4] を採用した。

統計分析において採用した ILR 変換は、組成データをユークリッド空間へ等長写像する正規直交座標を与えるものである。これにより、ユークリッド空間と同じ幾何学的性質に基づいて距離や分散を定義でき、通常の統計手法を適用することが可能となる。しかし、ILR 座標は成分群間の対数比として構成されるため、各座標が単一成分の効果を直接表すものではなく、SHAP による特徴量寄与の解釈を難しくする。一方、CLR 変換は、変換前後で次元数が変わらず、各成分の対数比を直接表現できるため、SHAP による寄与方向の解釈が比較的明瞭となる。ただし、CLR 変換後のデータは総和ゼロ制約を有し、ユークリッド空間における独立な正規直交座標とはならないという理論的制約がある。本分析の機械学習において採用した決定木系アルゴリズム（XGBoost）は共分散構造や行列の正則性に依存しない学習機構を有することから、CLR 変換に伴う理論的制約の影響は限定的であると考えられる。以上より、本分析では、2 種類の変換方法（ILR 変換、CLR 変換）を分析目的に応じて使い分けることとした。

分析は、次の手順により行った。

（1）ゼロ置換（対数比変換の前処理）

（2）CLR（Centered Log-Ratio）変換

（3）機械学習モデルの作成

（4）予測モデルによる予測値の算出

（5）予測モデルの解釈および要因分析

具体的な方法を以下に示す。

（1）ゼロ置換（対数比変換の前処理）

「3.7.4　用途構成」の「(1) ゼロ置換（対数比変換の前処理)」に同じ。

（2）CLR（Centered Log-Ratio）変換

組成データを CLR 変換（中心対数比変換）により単体空間から実空間に写像した。CLR 変換は、各成分を全成分の幾何平均との対数比として表す変換であり、D 次元の組成ベクトルを同じく、D 次元の実数ベクトルとして表現する。ただし、変換後のベクトルは総和ゼロ制約を満たすため、$D-1$ 次元の部分空間上に存在する。

CLR 変換の第 i 成分 $clr_i(\mathbf{x})$ は、各成分を全成分の幾何平均との対数比として定義される中心対数比座標であり、式（18）のように表される。

$$clr_i(\boldsymbol{x}) = log\frac{x_i}{\left(\prod_{j=1}^{D} x_j\right)^{1/D}} \qquad (i = 1,\quad 2,\quad \ldots,\quad D) \tag{18}$$

ここで、

D　　：組成ベクトルの成分の数

$clr(\ \ )$　：CLR変換を表す関数

$\mathbf{x}$　　：組成ベクトル

x_i　　：組成ベクトル $\mathbf{x}$ の成分

添え字 i　：用途（1：暖房、2：冷房、3：給湯、4：調理、5：照明他）

である。この変換により、定数和制約は形式的には残るものの、各成分の相対的な寄与を直接評価できる表現となるため、木系機械学習モデルにおける特徴量解釈が明瞭となる。

（3）機械学習による予測モデルの作成

機械学習による予測モデルの作成は、「3.8.3　年間一次エネルギー消費量」の「(1) 機械学習による予測モデルの作成」に同じ。

5 分割交差検証によるモデルの評価結果およびハイパーパラメータの調整結果をそれぞれ表 3.12 および表 3.13 に示す。決定したハイパーパラメータを用いて、全分析対象世帯を対象に再学習を行った。再学習により得られたモデルを最終的な予測モデルとした。

表 3.12　5 分割交差検証によるモデルの評価結果

指標	単位	暖房比率		冷房比率		給湯比率		調理比率		照明他比率	
		平均	標準偏差	平均	標準偏差	平均	標準偏差	平均	標準偏差	平均	標準偏差
MAE（平均絶対誤差）	GJ/年	0.80	0.011	1.61	0.021	0.60	0.006	0.57	0.008	0.55	0.005
MSE（平均二乗誤差）	$GJ^2/年^2$	1.71	0.070	6.23	0.101	0.70	0.011	0.72	0.031	0.60	0.008
RMSE（平均二乗誤差の平方根）	GJ/年	1.31	0.027	2.50	0.020	0.83	0.006	0.85	0.018	0.78	0.005
R^2（決定係数）	－	0.73	0.011	0.58	0.006	0.45	0.013	0.61	0.020	0.44	0.009
RMSLE（平均二乗対数誤差の平方根）	－	0.41	0.007	0.52	0.004	0.33	0.003	0.31	0.004	0.23	0.002
MAPE（平均絶対パーセント誤差）	－	3.27	0.416	14.56	15.886	2.10	0.391	1.18	0.137	0.31	0.018

表 3.13　ハイパーパラメータの調整結果

変数名	変数名（日本語）	値（調整後）				
		暖房比率	冷房比率	給湯比率	調理比率	照明他比率
booster	ブースティング方式	gbtree				
tree_method	木構築方式	auto				
grow_policy	木成長方針	depthwise				
sampling_method	サンプリング方式	uniform				
n_estimators	木の本数	150	270	210	140	210
learning_rate	学習率	0.10	0.05	0.10	0.05	0.10
max_depth	木の最大深さ	7	6	6	11	6
min_child_weight	最小子ノード重み	3	2	4	3	4
subsample	行サンプリング率	0.7	0.7	1.0	0.5	1.0
colsample_bytree	列サンプリング率	1.0	0.9	0.9	1.0	0.9
gamma	分割最小損失減少量	0	0	0	0	0
reg_alpha	L1 正則化係数	5.00	0.00	3.00	0.01	3.00
reg_lambda	L2 正則化係数	10.000	0.000	2.000	4.000	2.000
scale_pos_weight	クラス重み調整係数	2.2	1.6	10.2	12.2	10.2
base_score	初期予測値	0.30	-3.25	1.24	-0.50	2.22
objective	目的関数	reg:squarederror				
eval_metric	評価指標	rmse				
random_state	乱数シード	1	1	1	1	1
seed	学習用シード値	1	1	1	1	1
seed_per_iteration	反復ごとのシード設定	0	0	0	0	0

(4) 予測モデルによる予測値の算出

「3.8.3　年間一次エネルギー消費量」の「(2) 予測モデルによる予測値の算出」に同じ。

(5) 予測モデルの解釈および要因分析

「3.8.3　年間一次エネルギー消費量」の「(3) 予測モデルの解釈および要因分析」に同じ。

ただし、機械学習モデル作成ではCLR変換した構成比率を目的変数としていることから、SHAP値は構成比率そのものの増減を直接的に示すものではない。また、CLR 変換は定数和制約を緩和する一方で、用途間の相対的な依存関係を保持する変換である。そのため、個々の特徴量に対する SHAP 値を構成比率の変化量として解釈することは適切ではない。以上を踏まえ、本研究では特徴量重要度および要因別の平均 SHAP 値に基づく分析に限定し、特徴量ごとの SHAP 値の分布に基づく分析を省略する。

3.9 使用したプログラミング言語環境およびライブラリ・パッケージ名

気象データの作成、特徴量の作成、集計、統計分析、機械学習、モデル解釈とデータ可視化は、表 3.14 に示す、Python および R のライブラリ・パッケージを使用することで行った。

表 3.14 使用したイブラリ・パッケージ名の一覧 その 1

(a) 気象データの作成 1
(「ArcClimate」[10] による任意の地点座標における毎時の気象データの作成)

用途	開発言語	ライブラリ・パッケージ名
プログラミング言語環境	Python	Python 3.13.0
データ分析・データ操作	Python	Pandas 2.2.3
数値計算・配列演算	Python	NumPy 2.1.3

(b) 気象データの作成 2
(「grib2-to-csv」[11] によるデータ形式変換)

用途	開発言語	ライブラリ・パッケージ名
プログラミング言語環境	Python	Python 3.9.16
データ分析・データ操作	Python	Pandas 2.1.4
数値計算・配列演算	Python	NumPy 1.26.3
統計分析・数値最適化	Python	SciPy 1.11.4
地理空間データ処理 (ラスタ・ベクタ変換等)	Python	GDAL 3.6.2

(c) 特徴量の作成・集計・統計分析

用途	開発言語	ライブラリ・パッケージ名
プログラミング言語環境	Python	Python 3.13.0
データ分析・データ操作	Python	Pandas 2.2.3
数値計算・配列演算	Python	NumPy 2.1.3
統計分析・統計モデリング (多重比較補正 : Holm 法等)	Python	statsmodels 0.14.4
統計分析・数値最適化	Python	scipy 1.15.3
組成データ解析 (対数比変換、PERMANOVA)	Python	scikit-bio 0.7.1
Python-R 連携	Python	rpy2 3.6.4
プログラミング言語環境	R	R 4.3.3
データ分析・基本統計	R	base 4.3.3
統計分析 (回帰、検定等)	R	stats 4.3.3
統計分析 (ノンパラメトリック多重比較 : Steel-Dwass 法)	R	NSM3 1.20
組成データ解析 (Aitchison 平均)	R	compositions 2.0.8
組成データ解析 (ゼロ値置換)	R	zCompositions 1.5.0.5
多変量解析 (PERMANOVA 等)	R	vegan 2.6.10
データ可視化	Python	seaborn 0.13.2
データ可視化	Python	Matplotlib 3.10.0

表 3.10　使用したイブラリ・パッケージ名の一覧　その 2

（d）機械学習・モデル解釈

用途	開発言語	ライブラリ・パッケージ名
プログラミング言語環境	Python	Python 3.11.14
データ分析・データ操作	Python	Pandas 2.1.4
数値計算・配列演算	Python	NumPy 1.26.4
機械学習（汎用）	Python	scikit-learn 1.4.2
自動機械学習（AutoML）	Python	PyCaret 3.3.2
機械学習（勾配ブースティング）	Python	XGBoost 2.0.3
モデル解釈・説明可能性分析	Python	SHAP 0.48.0
データ可視化	Python	seaborn 0.13.2