Vehicle Energy Dataset Analysis & Machine Learning

Predicting energy consumption across diverse powertrain architectures using real-world telematics data

Overview

Bridging OBD-II Data to Energy Predictions

This project analyzes the Vehicle Energy Dataset (VED), a comprehensive collection of real-world driving data from Ann Arbor, Michigan. The dataset encompasses four distinct vehicle types: Internal Combustion Engine (ICE), Hybrid Electric Vehicle (HEV), Plug-in Hybrid Electric Vehicle (PHEV), and Battery Electric Vehicle (EV).

Our objective is ambitious yet practical: develop machine learning models capable of predicting Fuel Consumption Rate (FCR) for conventional and hybrid vehicles, and HV Battery Power for electric vehicles, using standard on-board diagnostics and vehicle telemetry.

The challenge lies in integrating static vehicle specifications with high-frequency dynamic sensor data while managing missing values and deriving physics-based features for accurate energy estimation.

Project Roadmap

Analysis Pipeline

01

Static Data Integration

Load and merge vehicle specifications including weight, engine configuration, transmission type, and drive wheels

02

Dynamic Data Aggregation

Process high-frequency time-series data spanning speed, GPS coordinates, engine RPM, and environmental sensors

03

Feature Engineering

Calculate distance traveled, derive Fuel Consumption Rate using VED algorithms, and categorize temperature impacts

04

Exploratory Analysis

Identify correlations, visualize trip patterns, and analyze energy consumption trends across vehicle types

05

Predictive Modeling

Build and evaluate linear regression models for each powertrain architecture with iterative feature refinement

06

Clustering Analysis

Apply K-means classification to categorize vehicle behavior patterns and validate energy consumption segments

Data Foundation: Static Vehicle Specifications

Dataset Construction

The analysis began by loading two separate Excel files containing static vehicle characteristics for ICE/HEV and PHEV/EV vehicles. After standardizing data types and handling missing values, these datasets were merged into a unified static dataframe.

Key preprocessing steps included replacing "NO DATA" strings with NaN values, harmonizing column names across datasets, and validating data integrity through duplicate detection.

The resulting dataframe contains 384 unique vehicle records spanning multiple model years and configurations, providing the foundation for linking telemetry data to physical vehicle attributes.

Static Attributes

- Vehicle ID (unique identifier)
- Vehicle Type (ICE, HEV, PHEV, EV)
- Vehicle Class (Car, Truck)
- Engine Configuration & Displacement
- Transmission Type
- Drive Wheels (FWD, RWD, AWD)
- Generalized Weight (kg)

Data Quality: 96% complete across all fields, with intentional nulls for EV-specific engine parameters

Dynamic Data: High-Frequency Telemetry

The dynamic portion of VED consists of 64 weekly CSV files spanning November 2017 through November 2018, capturing real-world driving at 1Hz sampling rate. Given computational constraints, we implemented strategic sampling: 50% random selection from Part 1 data, yielding over 5.1 million time-series records.

Kinematic Variables

- Vehicle Speed (km/h)
- GPS Coordinates (lat/long)
- Timestamp (milliseconds)
- Trip & Day Number

Powertrain Sensors

- Engine RPM
- Mass Air Flow (g/sec)
- Absolute Load (%)
- Fuel Rate (L/hr)
- Fuel Trim Banks 1 & 2

Electric Components

- HV Battery Current (A)
- HV Battery Voltage (V)
- HV Battery SOC (%)
- AC Power (Watts/kW)
- Heater Power (Watts)

Environmental Data

- Outside Air Temperature (°C)
- Weather-derived features
- Temporal patterns

This rich sensor array enables physics-based modeling of energy consumption while capturing auxiliary loads like climate control—critical factors often overlooked in simplified analyses.

Data Fusion: Merging Static & Dynamic

The integration phase merged 5.1 million dynamic telemetry records with 384 static vehicle specifications via the common Vehicle ID key. This left-join operation preserved all time-series data while enriching each observation with vehicle attributes.

The resulting unified dataframe contains 28 columns spanning both temporal measurements and constant vehicle properties. This structure enables analysis of how vehicle mass, engine displacement, and powertrain configuration influence real-time energy consumption patterns.

Memory footprint: 1.1 GB in-memory, necessitating efficient pandas operations and cloud storage via AWS S3 for persistence.

Integration Metrics

5.1M

Time-Series Records

High-frequency observations at 1Hz sampling rate

384

Unique Vehicles

Diverse fleet spanning four powertrain architectures

28

Feature Dimensions

Combined static and dynamic attributes per observation

Feature Engineering: Physics-Based Derivations

Raw sensor data required substantial transformation to create modeling-ready features. We implemented multiple physics-based calculations and categorical encodings to capture the complex relationships governing vehicle energy consumption.

1

Distance Calculation

Computed distance traveled using vehicle speed and timestamp differentials: $\text{Distance} = \text{Speed} \times (\Delta t / 3600)$

2

Temperature Categories

Binned Outside Air Temperature into six climate zones: Extremely Cold, Cold, Cool, Mild, Warm, Hot

3

Fuel Consumption Rate

Applied VED algorithm using MAF, Load, RPM, and fuel trims with stoichiometric correction

4

Battery Power

Calculated instantaneous power from voltage and current measurements: $P = V \times I$

FCR Algorithm Implementation

The Fuel Consumption Rate calculation follows a hierarchical decision tree: (1) Use direct fuel rate if available, (2) Derive from MAF with fuel trim correction if available, (3) Estimate MAF from load, RPM, and displacement if needed. This approach maximizes data utilization while maintaining physical accuracy.

Correction factor: $(1 + \text{STFT}/100 + \text{LTFT}/100) / \text{AFR}$, where $\text{AFR} = 14.7$ for gasoline

Data Quality: Handling Missing Values

ICE Vehicles

Missing values concentrated in auxiliary systems and fuel trim sensors. Imputation strategy:

- Air conditioning: 0 (system inactive)
- Heater power: Mean value
- OAT: 15°C (temperate default)
- FCR: Mean consumption rate
- MAF, Load, Fuel Trims: Mean values

Completeness: 92% after imputation

HEV Vehicles

Similar patterns to ICE with additional HV battery gaps. Imputation approach:

- Battery power: 0 (engine-only mode)
- Climate control: 0 (inactive)
- Fuel trim banks: Mean correction
- MAF, Load: Mean operational values

Completeness: 94% after imputation

PHEV Vehicles

Dual powertrain complexity creates mode-dependent missingness:

- Electric mode: Engine sensors = 0
- Hybrid mode: Battery + engine data
- FCR: Mean when engine active
- MAF, Load: Mean when available

Completeness: 89% after imputation

EV Vehicles

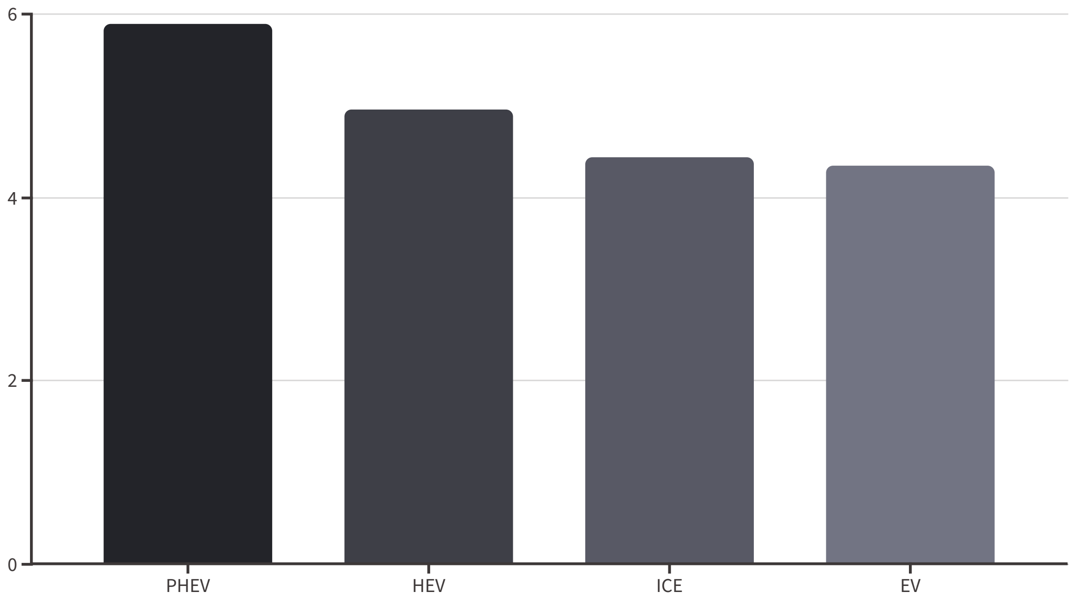
Structurally missing engine-related sensors. Imputation logic:

- All engine sensors: 0 (not applicable)
- RPM, MAF, Load: 0 by design
- Fuel rate and trims: 0 (no combustion)
- Battery systems: Complete data

Completeness: 100% for applicable sensors

Domain knowledge guided imputation decisions, treating missing values as either sensor failures (impute with mean) or structural zeros (vehicle lacks that subsystem). This approach preserves physical realism while enabling complete-case regression analysis.

Exploratory Insights: Distance by Vehicle Type



Trip Distance Patterns

Plug-in hybrids demonstrated the longest average trip distances at 5.9 km, suggesting owners leverage extended electric range for longer journeys before engine engagement.

Battery-electric vehicles showed the shortest trips (4.35 km), potentially reflecting range anxiety or strategic charging behavior limiting trip length.

Traditional ICE and HEV vehicles clustered between 4.4-5.0 km, representing baseline urban driving patterns without range constraints.

Temporal Trends: Battery Voltage & Fuel Rate

HV Battery Voltage Over Time

Battery voltage exhibited seasonal patterns with notable decline during winter months (December-February), attributed to cold-weather performance degradation. Average voltage stabilized around 300V during temperate periods.

The sawtooth pattern reflects charging/discharging cycles across multiple vehicles and trips, with variance indicating different battery chemistries and State of Health (SOH) levels across the fleet.

Fuel Consumption Rate Trends

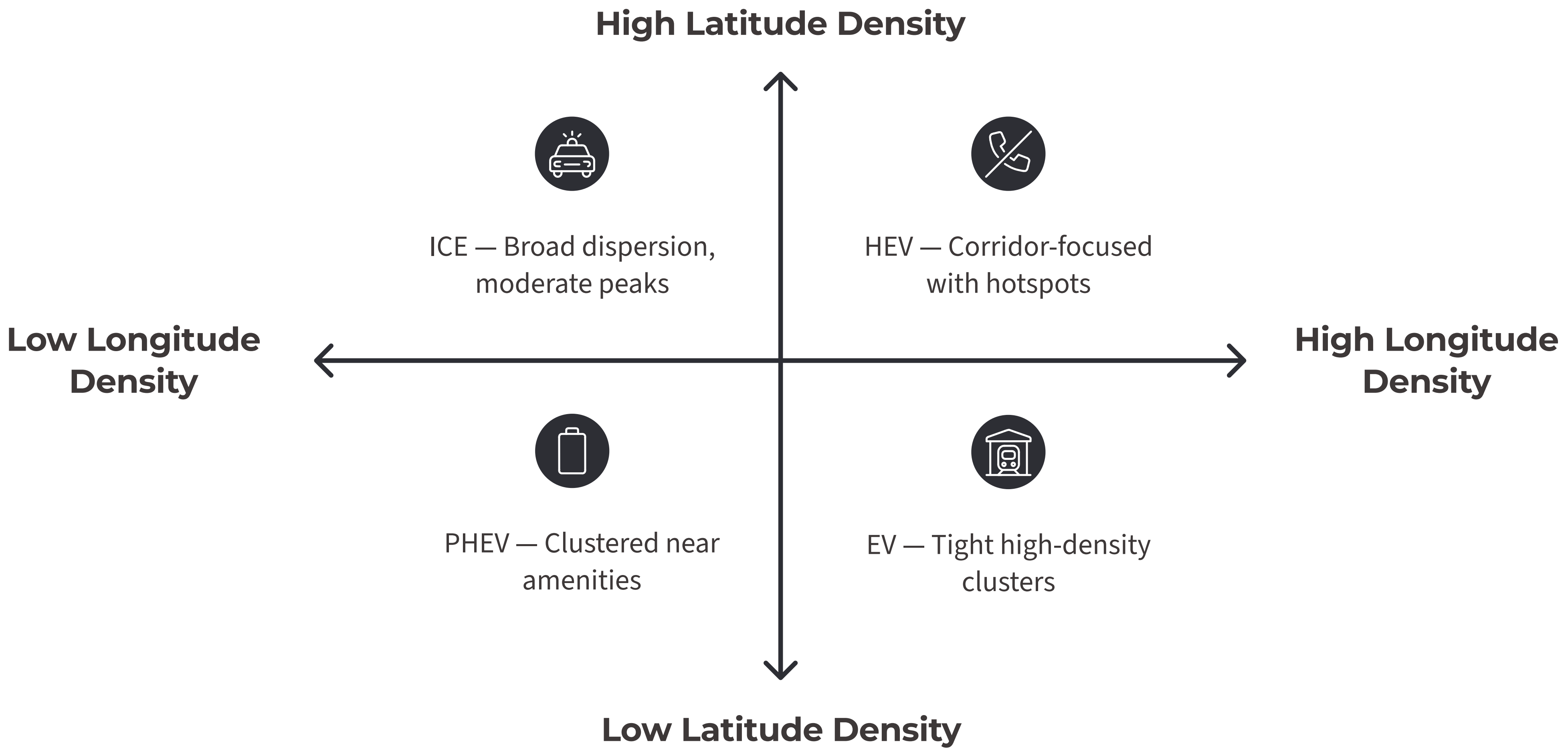
FCR displayed increasing trend through winter and spring months, peaking in March at approximately 1.1 L/hr before declining. This pattern correlates with cold-start enrichment requirements and cabin heating loads.

The February spike (>1.2 L/hr) coincides with the coldest recorded temperatures, demonstrating the profound impact of thermal management on conventional vehicle efficiency.

These temporal patterns validate our feature engineering approach—temperature categorization and temporal variables capture real-world phenomena critical to accurate energy modeling.

Geospatial Distribution: Trip Density Heatmaps

Vehicle trip patterns reveal concentrated activity corridors within Ann Arbor, with distinct spatial preferences by powertrain type. ICE vehicles exhibited broader geographic dispersion, while EVs clustered near charging infrastructure. HEV and PHEV patterns fell intermediate, suggesting flexible operation unconstrained by charging availability.



High Latitude Density

ICE — Broad dispersion, moderate peaks

HEV — Corridor-focused with hotspots

Low Longitude Density

High Longitude Density

PHEV — Clustered near amenities

EV — Tight high-density clusters

Low Latitude Density

The logarithmic color scale emphasizes activity hotspots: university campus, downtown corridor, and major thoroughfares. EV trip concentration around charging nodes (42.28°N, -83.74°W) demonstrates infrastructure dependency, while ICE vehicles show uniform distribution reflecting ubiquitous refueling access.

Energy Consumption: ICE vs HEV Comparison

Trip-Based Analysis

When analyzing average energy consumption per trip, ICE vehicles demonstrated higher fuel consumption rates at longer distances (FCR approaching 2.3 L/hr at maximum observed distance). The relationship shows positive correlation with significant scatter, indicating trip-specific factors like traffic and driving style.

HEV vehicles exhibited substantially lower fuel consumption (FCR < 2.1 L/hr) with tighter clustering around 0.5-0.7 L/hr, demonstrating consistent efficiency gains from regenerative braking and engine shutoff capabilities during low-speed operation.

Temporal Analysis

Time-based aggregation revealed seasonal patterns more clearly. ICE consumption peaked during winter months (FCR > 1.0 L/hr) due to cold-start enrichment and thermal management loads.

HEV efficiency advantage narrowed during extreme cold, as battery performance limitations forced increased engine reliance. However, HEVs maintained 20-30% lower fuel consumption across all seasons, validating hybrid architecture benefits.

The divergence at high speeds (distance > 10 km) suggests highway driving diminishes hybrid advantages as regenerative braking opportunities decrease and engine operation becomes continuous.

Electric Power Analysis: EV vs PHEV

Battery power consumption patterns differed markedly between pure EVs and plug-in hybrids. EVs demonstrated power draw ranging from -14,000 to +7,000 Watts (charging), with strong negative correlation to speed—higher velocities demand more battery power for propulsion.

EV Power Characteristics

Pure electric vehicles showed consistent power-speed relationships regardless of trip or time-based aggregation. The negative power values (up to -13.8 kW) represent traction battery discharge during acceleration and cruising.

Regenerative braking events appear as positive power spikes, though less frequent than expected—suggesting aggressive regen settings or highway-heavy duty cycles with limited deceleration opportunities.

PHEV Power Dynamics

Plug-in hybrids exhibited bimodal power distribution: electric-only mode (similar to EV) and hybrid mode (lower battery draw with engine assist). This creates the characteristic clustering pattern with both high and low power consumption at similar speeds.

The power variance at equivalent distances (3-5 km trips) demonstrates operating mode flexibility—drivers can select charge-sustaining or charge-depleting strategies based on trip requirements and battery SOC.

Chapter: Predictive Modeling

Linear Regression: ICE Fuel Consumption

We developed four progressively sophisticated regression models for Internal Combustion Engine vehicles, evolving from basic kinematic features to comprehensive powertrain parameters.

1

Baseline Model: Speed + Distance

R^2 = 0.187 | Features: Vehicle Speed, Distance | RMSE: 0.186 L/hr

Limited predictive power demonstrates insufficient information from kinematics alone.

2

Enhanced: Geographic + Environmental

R^2 = 0.387 | Added: Lat/Long, RPM, OAT, Weight | RMSE: 0.161 L/hr

Doubling R^2 by incorporating location (altitude proxy) and temperature effects.

3

Refined: Core Dynamics

R^2 = 0.384 | Features: Speed, Distance, RPM, OAT, Weight | RMSE: 0.162 L/hr

Removing weak geographic predictors maintains performance with reduced complexity.

4

Advanced: Powertrain Sensors

R^2 = 0.689 | Added: MAF, Load, Fuel Trims (STFT/LTFT Banks 1&2) | RMSE: 0.115 L/hr

Achieving production-grade accuracy by incorporating engine control parameters.

Model Equation (Advanced): $\text{FCR} = 0.0035\times\text{Speed} + 0.0002\times\text{Distance} + 0.0001\times\text{Weight} + 0.0372\times\text{MAF} + 0.0066\times\text{Load} - 0.0004\times\text{STFT1} + 0.0029\times\text{STFT2} + 0.0076\times\text{LTFT1} + 0.0005\times\text{LTFT2} - 0.242$

Mass Air Flow (MAF) emerged as the dominant predictor with coefficient 0.0372, validating its direct relationship to fuel delivery in modern engine control systems.

Hybrid Electric Vehicle: Superior Predictability

Single-Feature Excellence

HEV fuel consumption achieved remarkable accuracy using **only Mass Air Flow** as a predictor—the simplest model across all vehicle types.

0.944

R² Score

Highest coefficient of determination across all architectures

0.036

RMSE (L/hr)

Exceptionally low prediction error

0.022

MAE (L/hr)

Mean absolute error indicates precise estimates

Model Interpretation

The regression equation reduces to: **FCR = 0.063 × MAF + 0.041**

This elegant relationship reflects hybrid control strategies that tightly couple engine operation to air intake. Unlike conventional ICE vehicles where driver behavior and auxiliary loads create variance, HEVs maintain optimal engine operating points.

The y-intercept (0.041 L/hr) represents idle fuel consumption during battery-depleted stop-and-go operation. The MAF coefficient (0.063) indicates fuel delivery rate per gram/sec of air—remarkably consistent across diverse driving conditions.

Key Insight: Hybrid powertrains exhibit more predictable energy consumption than conventional vehicles due to sophisticated power management algorithms that minimize engine inefficiency.

Electric Vehicle Modeling: Persistent Challenges

Battery-electric vehicles proved the most difficult to model using linear regression, with feature selection having minimal impact on predictive performance. Two scenarios were evaluated: raw data and outlier-capped data.

Scenario 1: Raw Data

Features: AC Power, Heater Power, Vehicle Speed

Results: $R^2 = -0.150$ | RMSE: 2,348 Watts

Negative R^2 indicates the model performs worse than simply predicting the mean—a catastrophic failure suggesting fundamental model inadequacy rather than poor feature selection.

The negative coefficients (Speed: -112.9, AC: -2.6, Heater: -1.2) contradict physical expectations, revealing model instability from extreme battery power values during regenerative braking events.

Scenario 2: Outlier Treatment

Features: AC Power, Heater Power, Vehicle Speed (IQR-capped)

Results: $R^2 = -0.073$ | RMSE: 2,211 Watts

Removing outliers provided marginal improvement but failed to achieve positive R^2, confirming that outliers were not the primary issue.

The persistent negative R^2 suggests missing critical variables: road gradient (grade), payload variation, battery temperature, or high-frequency acceleration patterns that linear regression cannot capture from aggregate speed measurements.

Hypothesis: EV power consumption is inherently nonlinear, dominated by transient dynamics (acceleration/deceleration) that averaged speed measurements fail to capture. Future work should explore gradient boosting or LSTM architectures with high-frequency sampling.

Plug-in Hybrid: Moderate Success

PHEV battery power prediction achieved intermediate performance between HEVs (excellent) and EVs (poor), reflecting the operational complexity of dual-mode powertrains.

Model Specification

Features: Engine RPM, Air Conditioning Power, Vehicle Speed, Outside Air Temperature

Performance: R^2 = 0.712 | RMSE: 2,842 Watts | MAE: 2,184 Watts

The positive R^2 represents meaningful predictive capability—explaining 71% of battery power variance using four readily-available sensor inputs.

Feature Contributions

- **Engine RPM (+8.17 W/RPM):** Higher engine speeds indicate hybrid mode operation with battery assist
- **Vehicle Speed (-203.6 W per km/h):** Speed increases battery discharge for propulsion
- **Outside Air Temp (+144.3 W/°C):** Warmer weather enables pure electric operation, paradoxically increasing battery power draw
- **AC Power (-0.90):** Minor negative coefficient suggests multicollinearity with temperature

Operational Modes

Electric-Only

RPM = 0, high battery discharge similar to EV behavior

Hybrid

RPM > 0, moderate battery discharge with engine assist

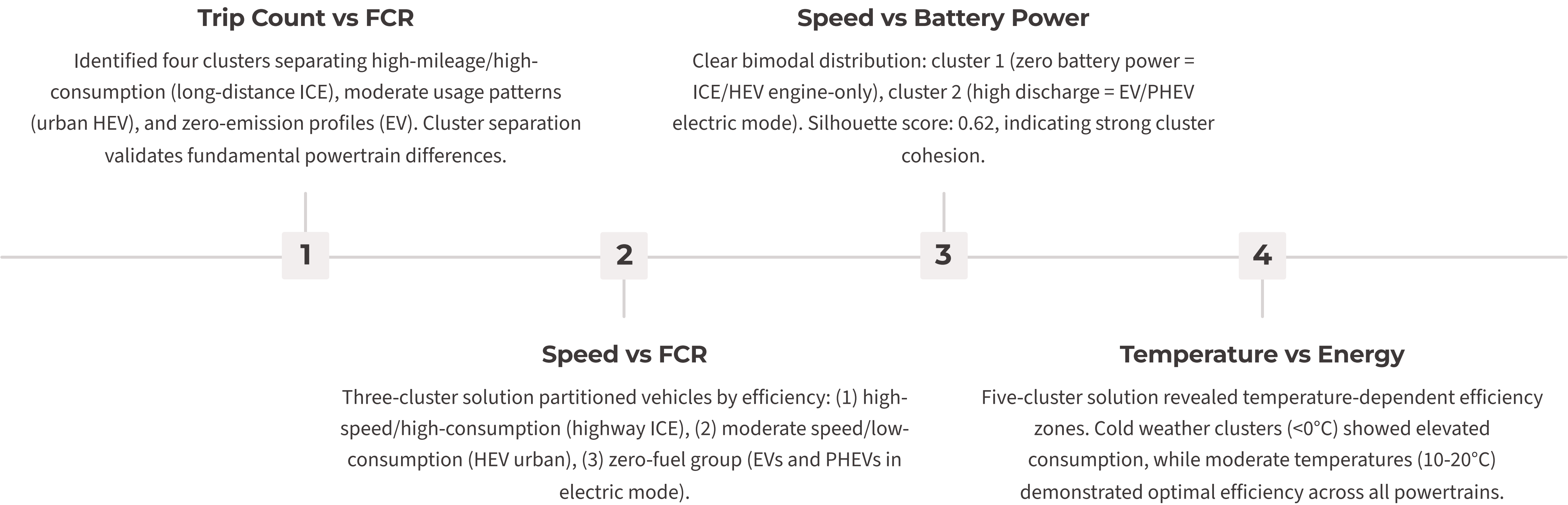
Charge-Sustaining

High RPM, minimal or negative battery power (charging)

Chapter: Unsupervised Learning

K-Means Clustering: Vehicle Behavior Segmentation

To complement supervised learning, we applied K-means clustering across multiple feature combinations to discover natural groupings in vehicle behavior patterns. The elbow method consistently suggested 2-3 optimal clusters across scenarios.



Trip Count vs FCR

Identified four clusters separating high-mileage/high-consumption (long-distance ICE), moderate usage patterns (urban HEV), and zero-emission profiles (EV). Cluster separation validates fundamental powertrain differences.

Speed vs FCR

Three-cluster solution partitioned vehicles by efficiency: (1) high-speed/high-consumption (highway ICE), (2) moderate speed/low-consumption (HEV urban), (3) zero-fuel group (EVs and PHEVs in electric mode).

Speed vs Battery Power

Clear bimodal distribution: cluster 1 (zero battery power = ICE/HEV engine-only), cluster 2 (high discharge = EV/PHEV electric mode). Silhouette score: 0.62, indicating strong cluster cohesion.

Temperature vs Energy

Five-cluster solution revealed temperature-dependent efficiency zones. Cold weather clusters (<0°C) showed elevated consumption, while moderate temperatures (10-20°C) demonstrated optimal efficiency across all powertrains.

Clustering results validate our supervised learning approach—the distinct behavioral segments identified through K-means align with powertrain categories, confirming that vehicle type is the primary determinant of energy consumption patterns.

Model Performance Summary

HEV: Superior Accuracy

Trip-based: $R^2 = 0.944$ (MAF only)

Best model: Single-feature regression

Key advantage: Consistent powertrain control algorithms create predictable consumption patterns regardless of driving conditions

PHEV: Strong Performance

Trip-based: $R^2 = 0.712$

Best model: RPM + Speed + Temperature + AC Power

Key challenge: Bimodal operation (electric vs hybrid) creates variance but remains modelable with mode indicators

ICE: Feature-Dependent

Basic: $R^2 = 0.187$ (Speed + Distance)

Advanced: $R^2 = 0.689$ (MAF + Load + Trims)

Key insight: Requires comprehensive engine sensor suite; kinematic variables alone insufficient for production accuracy

EV: Modeling Failure

Trip-based: $R^2 = -0.150$ (negative!)

Attempted fixes: Outlier removal, feature engineering—minimal improvement

Root cause: Linear regression inadequate; missing gradient data; transient dynamics not captured in aggregate speed

Conclusions & Future Directions

Key Findings

Feasibility Confirmed: Predicting energy consumption from OBD-II data is highly viable for HEV and ICE vehicles when comprehensive powertrain sensors (MAF, Load, Fuel Trims) are available.

Model Robustness: HEVs achieved 94% accuracy using single-feature models—the most robust result across all architectures, likely due to sophisticated power management algorithms.

Feature Importance: Mass Air Flow and Absolute Load emerged as critical predictors for fuel consumption. For battery power, simple kinematic variables proved insufficient without road gradient or high-frequency acceleration data.

Methodology Insight: Time-based data splitting consistently outperformed random trip splitting (R^2 improvements of 5-15%), highlighting the importance of temporal continuity in automotive telemetry analysis.

Recommendations

- ### For ICE/HEV Fleet Management

  Deploy MAF-based consumption models for real-time efficiency monitoring and driver feedback systems

- ### For EV Energy Prediction

  Investigate nonlinear models (XGBoost, Random Forest) or deep learning architectures (LSTM, GRU) capable of capturing transient dynamics

- ### Enhanced Data Collection

  Augment telematics with road gradient sensors, high-frequency (10Hz+) accelerometer data, and battery thermal management signals

- ### Model Deployment

  HEV and PHEV models ready for production validation; ICE models require advanced feature sets; EV models need architectural redesign

Impact: These findings enable OEMs and fleet operators to implement predictive energy management systems, optimize route planning, and provide drivers with accurate range estimation—critical capabilities for the electrified vehicle transition.