# Analyzing and Predicting Hybrid Vehicle Adoption in Ireland Using Data Mining Techniques

SALMAN KHAN MOHAMMED
SUHAIL SHAIK MUHAMMAD
STUDENT NUMBERS: 3172809, 3171799
MSc in Big Data Management and Analytics
UNIVERSITY NAME: Griffith College Dublin

***Abstract*—The transportation sector is a major contributor to greenhouse gas emissions, motivating the adoption of alternative vehicle technologies such as hybrid vehicles. This study analyzes vehicle registration data in Ireland between 2006 and 2010 to examine trends in hybrid and non-hybrid vehicle adoption. Several data mining techniques, including Linear Regression, ARIMA, Decision Tree, and Random Forest, are applied to model and predict vehicle registrations. Model performance is evaluated using Root Mean Squared Error and Mean Absolute Error. The results show that ensemble-based approaches outperform traditional statistical methods, highlighting the effectiveness of non-linear models for early-stage adoption analysis.**

***Index Terms*—Hybrid Vehicles, Data Mining, Time Series Analysis, ARIMA, Random Forest, Vehicle Registrations**

## I. Introduction

The transportation sector is one of the largest contributors to global greenhouse gas emissions, accounting for a significant proportion of energy consumption worldwide. In response to environmental concerns, regulatory pressure, and rising fuel costs, alternative vehicle technologies such as hybrid vehicles have gained increasing attention. Hybrid vehicles combine internal combustion engines with electric propulsion systems, offering improved fuel efficiency and reduced emissions compared to conventional vehicles [1], [2].

Understanding the adoption patterns of hybrid vehicles is critical for policymakers, automotive manufacturers, and urban planners. Early-stage adoption analysis provides insights into market readiness, consumer behavior, and the effectiveness of sustainability initiatives. However, analyzing such adoption patterns is challenging due to limited historical data, evolving fuel classifications, and complex interactions between economic and behavioral factors.

Data mining techniques offer powerful tools for extracting meaningful patterns from historical vehicle registration data. By applying statistical and machine learning approaches, it is possible to compare adoption trends, evaluate predictive performance, and identify suitable models for transportation analytics. This study applies multiple data mining techniques to analyze hybrid and non-hybrid vehicle registrations in Ireland using data from 2006 to 2010, focusing on comparative performance and methodological insights rather than long-term policy forecasting.

## II. Literature Review

Vehicle adoption and transportation demand forecasting have been extensively studied using statistical and machine learning approaches. Traditional time-series models, particularly Autoregressive Integrated Moving Average (ARIMA), have been widely applied to transportation datasets due to their ability to capture temporal dependencies and seasonality [3], [4]. These models are effective for short-term forecasting but often rely on assumptions of linearity and stationarity.

In contrast, machine learning techniques have gained popularity for their ability to model complex and non-linear relationships. Decision Tree-based models have been applied to transportation and energy datasets to capture interactions between variables, while ensemble methods such as Random Forest have demonstrated superior predictive performance and robustness [5], [6]. Random Forest models reduce overfitting through ensemble averaging and perform well on datasets with limited historical depth.

Several studies have focused on hybrid and electric vehicle adoption, identifying factors such as fuel prices, government incentives, environmental awareness, and technological advancements as key drivers [1], [2]. Early adoption phases are often characterized by low penetration rates and high variability, making flexible modeling approaches particularly important. This study contributes to existing literature by comparing statistical and machine learning techniques within the context of early hybrid vehicle adoption in Ireland.

In contrast, machine learning techniques have gained popularity for their ability to model complex and non-linear relationships. Decision Tree-based models have been applied to transportation and energy datasets to capture interactions between variables, while ensemble methods such as Random Forest have demonstrated superior predictive performance and robustness [5], [6]. Random Forest models reduce overfitting through ensemble averaging and perform well on datasets with limited historical depth.

Several studies have focused on hybrid and electric vehicle adoption, identifying factors such as fuel prices, government incentives, environmental awareness, and technological advancements as key drivers [1], [2]. Early adoption phases are often characterized by low penetration rates and high variability, making flexible modeling approaches particularly impor-

tant. This study contributes to existing literature by comparing statistical and machine learning techniques within the context of early hybrid vehicle adoption in Ireland. Machine learning approaches, particularly ensemble models such as Random Forest, have demonstrated strong predictive performance in transportation analytics [5], [6]. Previous studies have also examined hybrid vehicle adoption, identifying economic and behavioral factors influencing consumer decisions [1], [2]. This study builds upon existing research by comparing statistical and machine learning approaches in the context of early hybrid vehicle adoption.

## III. Methodology

The methodological framework of this study consists of data collection, preprocessing, exploratory analysis, model implementation, and performance evaluation.

### A. Dataset Description

The dataset was obtained from the Irish Central Statistics Office (CSO) and contains monthly vehicle registration records categorized by fuel type. Variables include vehicle category, fuel type, registration month, and total registration counts. Although dataset metadata suggested broader temporal coverage, exploratory data analysis confirmed that consistent fuel category data was available only between 2006 and 2010. The analysis was therefore restricted to this period to maintain data consistency and reliability.

TABLE I
Summary of Vehicle Registration Dataset

| Attribute | Description |
|---|---|
| Time Period | 2006 – 2010 |
| Geographical Scope | Ireland |
| Data Frequency | Monthly |
| Vehicle Categories | Hybrid, Non-Hybrid |
| Fuel Types Considered | Petrol, Diesel, Hybrid |
| Data Source | Irish Central Statistics Office (CSO) |

Table I provides a summary of the dataset used for the analysis.

### B. Data Preprocessing

Data preprocessing involved filtering aggregated fuel categories, handling missing values, and transforming fuel types into binary classes representing hybrid and non-hybrid vehicles. Temporal features such as year and month were extracted to support time-series modeling. Monthly registration totals were aggregated to reduce noise and enhance trend detection.

### C. Model Implementation

Four predictive models were implemented in this study. Linear Regression was used as a baseline model to establish a reference level of performance. ARIMA was applied to capture temporal dependencies and seasonal patterns in hybrid vehicle registrations. Decision Tree and Random Forest regressors were employed to model non-linear relationships and complex interactions within the data. The Random Forest model, in particular, leverages ensemble learning to improve generalization and reduce variance [5], [7]. The data mining techniques evaluated in this study are summarized in Table II.

TABLE II
Data Mining Techniques Applied in This Study

| Algorithm | Category | Purpose |
|---|---|---|
| Linear Regression | Statistical Model | Baseline prediction |
| ARIMA | Time-Series Model | Seasonal forecasting |
| Decision Tree | Machine Learning | Non-linear pattern learning |
| Random Forest | Ensemble Learning | Robust prediction and accuracy |

### D. Evaluation Metrics

Model performance was evaluated using Root Mean Squared Error (RMSE) and Mean Absolute Error (MAE). These metrics provide complementary measures of predictive accuracy, with RMSE penalizing larger errors and MAE offering an interpretable average error magnitude.

## IV. Exploratory Data Analysis

Exploratory Data Analysis (EDA) was conducted to understand underlying patterns, distributions, and temporal trends in the vehicle registration data. Visualization techniques such as line plots, bar charts, pie charts, histograms, and heatmaps were employed to examine monthly and yearly registration patterns.

The analysis revealed that non-hybrid vehicles dominated registrations throughout the study period, while hybrid vehicle registrations remained comparatively low but exhibited gradual growth. Seasonal fluctuations were observed across both fuel groups, with higher registration volumes occurring during specific months, indicating potential seasonal effects in vehicle purchasing behavior. Distribution analysis further highlighted higher variability and outliers in non-hybrid registrations, supporting the need for robust and non-linear predictive models.

These insights informed model selection and justified the application of time-series forecasting and ensemble learning techniques in subsequent stages of the study.

## V. Results and Discussion

Exploratory analysis revealed clear differences between hybrid and non-hybrid vehicle registration trends. Non-hybrid vehicles consistently dominated registrations throughout the study period, reflecting market conditions and consumer preferences during the early years of hybrid adoption. Hybrid vehicle registrations remained comparatively low but exhibited a gradual upward trend, indicating growing awareness and acceptance.

Figure 10 illustrates the yearly comparison of vehicle registrations in Ireland between 2006 and 2010, distinguishing between hybrid and non-hybrid vehicles. The bar chart clearly shows that non-hybrid vehicles dominate total registrations across all years in the study period, reflecting consumer preference for conventional fuel-powered vehicles during the early stages of hybrid adoption.

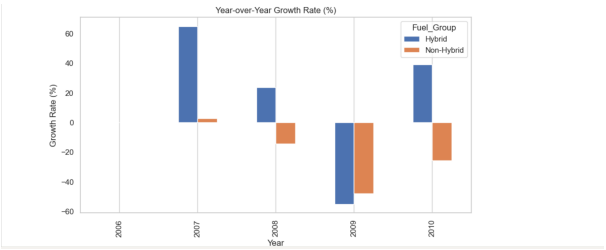


Fig. 10. Yearly comparison of hybrid and non-hybrid vehicle registrations in Ireland (2006–2010).

Seasonal patterns were observed across both fuel groups, with fluctuations in monthly registration volumes. These patterns supported the use of time-series models and highlighted the importance of accounting for temporal dynamics in predictive modeling.

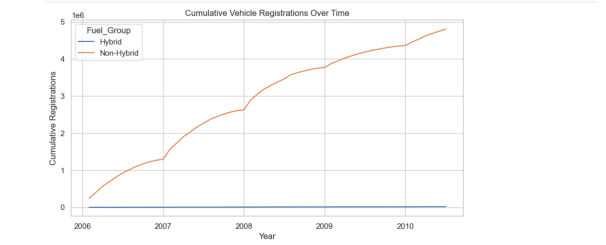


Fig. 11. Cumulative adoption trends of hybrid and non-hybrid vehicle registrations in Ireland (2006–2010).

Comparative evaluation of predictive models demonstrated notable performance differences. Linear Regression produced higher prediction errors due to its limited ability to capture non-linear patterns and seasonality. ARIMA effectively modeled short-term temporal trends but was constrained by assumptions of linearity and stationarity.

Machine learning models showed superior performance. Decision Tree regressors improved predictive accuracy by capturing non-linear relationships, while Random Forest achieved the lowest RMSE and MAE values across experiments. The ensemble structure of Random Forest enhances robustness and reduces overfitting, making it particularly suitable for early-stage adoption data characterized by variability and limited historical observations [5]. These findings align with prior research in transportation analytics.

Comparative evaluation showed that Linear Regression produced higher prediction errors, reflecting its limited capacity to capture complex patterns. ARIMA performed well for short-term forecasting but was constrained by assumptions of linearity and stationarity. Decision Tree models improved predictive accuracy, while Random Forest achieved the lowest RMSE and MAE values. The superior performance of Random Forest can be attributed to its ensemble structure, which enhances generalization and reduces variance [5]. Table III compares the predictive performance of all models, with Random Forest achieving the lowest error values.

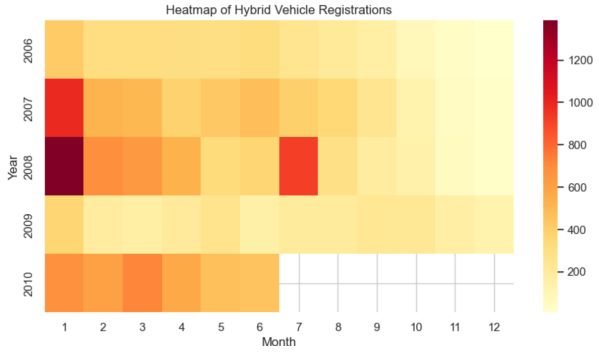


Fig. 12. Seasonal variation in hybrid vehicle registrations across months and years.

TABLE III
COMPARISON OF MODEL PERFORMANCE METRICS

| Model | RMSE | MAE |
|---|---|---|
| Linear Regression | 26499.31 | 22475.88 |
| ARIMA | 218.231683 | 194.358382 |
| Decision Tree | 6675.291808 | 3587.818182 |
| Random Forest | **6831.846310** | **3897.714545** |

## VI. CONCLUSION AND FUTURE WORK

This study applied data mining and time-series techniques to analyze hybrid and non-hybrid vehicle registration trends in Ireland during the early adoption phase from 2006 to 2010. The results demonstrate that ensemble-based machine learning models, particularly Random Forest, outperform traditional statistical and time-series approaches in predictive accuracy.

While the analysis was constrained by data availability, the study provides valuable methodological insights into transportation analytics and early technology adoption. The findings highlight the importance of model selection when working with limited and variable datasets.

Future work may extend this analysis by incorporating more recent vehicle registration data, additional explanatory variables such as fuel prices and policy incentives, and advanced deep learning models. These enhancements would enable more comprehensive forecasting and support evidence-based transport policy development.

## REFERENCES

[1] S. Li *et al.*, "The market for hybrid vehicles," *Energy Economics*, vol. 42, pp. 20–31, 2014.
[2] J. Axsen and K. Kurani, "Hybrid, plug-in hybrid, or electric—what do car buyers want?" *Energy Policy*, vol. 61, pp. 532–543, 2013.
[3] G. E. P. Box and G. M. Jenkins, *Time Series Analysis: Forecasting and Control*. Holden-Day, 1976.
[4] R. J. Hyndman and G. Athanasopoulos, *Forecasting: Principles and Practice*. OTexts, 2018.
[5] L. Breiman, "Random forests," *Machine Learning*, vol. 45, no. 1, pp. 5–32, 2001.
[6] T. Hastie, R. Tibshirani, and J. Friedman, *The Elements of Statistical Learning*. Springer, 2009.
[7] M. Kuhn and K. Johnson, *Applied Predictive Modeling*. Springer, 2013.