

---

# Unsupervised Temporal Clustering to Monitor the Performance of Alternative Fueling Infrastructure

---

Kalai Ramea<sup>1</sup>

## Abstract

Zero Emission Vehicles (ZEV) play an important role in the decarbonization of the transportation sector. For a wider adoption of ZEVs, providing a reliable infrastructure is critical. We present a machine learning approach that uses unsupervised temporal clustering algorithm along with survey analysis to determine infrastructure performance and reliability of alternative fuels. We illustrate this approach for the hydrogen fueling stations in California, but this can be generalized for other regions and fuels.

## 1. Introduction

The transportation sector accounts for about 33% of the total end-use carbon emissions highest among all the sectors in the US. Specifically, emissions from light-duty vehicles account for more than half of the total emissions in the transportation sector (AEO, 2018). Several countries have been adopting an increasing number of incentives to decarbonize the light-duty vehicle sector through zero-emission vehicles (ZEV). This has been reflected in the recent uptake of battery electric vehicle (BEV) fleet all over the world (Lutsey, 2015). However, there are still concerns among consumers about purchasing a BEV regarding range limitation, long charging times and charger congestion (Wahlman, 2013).

To further increase the total market penetration of ZEVs in the light-duty fleet, a portfolio approach of technological diversity is needed that addresses the variety of concerns expressed by the consumers. In California and Japan, hydrogen fuel cell vehicles (FCV) were introduced in recent years and are being increasingly adopted alongside BEVs. There are currently over 5,800 FCVs in California (CAFCP, 2019a), and it is estimated that about 40,000 FCVs will be on the Californian roads by 2022 (ARB, 2018). FCVs have a longer driving range, significantly shorter refueling time

<sup>1</sup>Palo Alto Research Center, Palo Alto, California, USA. Correspondence to: Kalai Ramea <kramea@parc.com>.

(in the order of minutes) and similar fueling mechanism as gasoline. Thus, they can play an important role in expanding the light-duty vehicle ZEV market as they have the potential to complement and alleviate some of the concerns raised about the BEVs.

One of the biggest barriers to FCV adoption is the reliability and availability of hydrogen fueling stations. Unlike BEVs, which have multiple modes of charging (home, work, public chargers), FCVs solely rely on access to public fueling infrastructure. The reputation of an unreliable station may discourage consumers from adopting the technology (Kelle, 2018). Therefore, for significant market acceptance of FCVs, policymakers should not only consider a good network of hydrogen refueling stations but also have a metric to assess their performance on a real-time basis.

**Existing Literature:** In the past, researchers have developed various designs of what constitutes an optimal network of hydrogen stations in the pre-planning stage (Ogden & Nicholas, 2011; Stephens-Romero et al., 2010). Several modeling approaches have been explored to understand the spread and quantity of hydrogen refueling station designs (Lim & Kuby, 2010; Nicholas & Ogden, 2007; Honma & Kurita, 2008; Bhatti et al., 2015; Kang & Recker, 2015). However, there has been little focus on assessing and monitoring the performance of existing stations. Although National Renewable Energy Laboratory has published specific data products on hydrogen station infrastructure (NREL, 2019), most of this data explains the technical details of the hydrogen fuel delivery and the adoption of fuel at the regional level on the supply side.

In this paper, we demonstrate an unsupervised temporal clustering approach on the hourly utilization data that was collected as a part of this project. To better understand the reasons behind the clustered categories, we conducted a survey of 100 FCV drivers in California. This approach identifies the stations that are performing within the healthy range, and those that are over-stressed and intervention is needed. We are publicly releasing the hourly station capacity dataset we collected for this research project<sup>1</sup>. To the

<sup>1</sup>Github link to the dataset will be released if the paper is accepted.

best of our knowledge, this is the first study that focuses on deploying a machine learning algorithm to analyze the micro-level usage of hydrogen fueling stations as a synchronous network to monitor their performance.

## 2. Methodology

The project was completed in three phases as follows: (a) Data collection, (b) Unsupervised temporal clustering, and (c) Survey analysis.

### 2.1. Data Collection and Pre-processing

To ground the analysis in real world data, we collected the hydrogen capacity data of all the fueling stations in California. The California Fuel Cell Partnership website ([CAFCP, 2019b](#)) provides information on the hydrogen fueling station locations and their real-time capacity levels in kg. The results shown in this paper is based on the hourly data collected for a period of three months starting from October 2018 to December 2018. A snapshot of utilization patterns of the stations is shown in Figure 1.

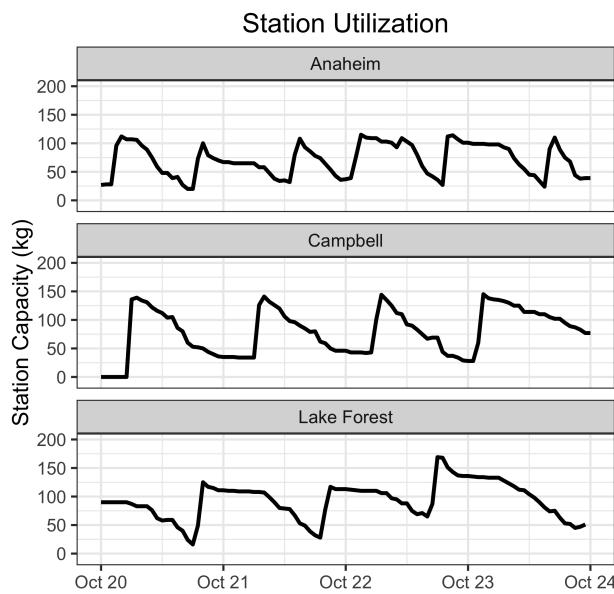


Figure 1. Snapshot of Hydrogen Fuel Utilization for Anaheim, Campbell and Lake Forest stations

The missing data in the time series due to technical issues in reporting were imputed through linear interpolation <sup>2</sup>. As the capacity varies across the stations, the time-series data

<sup>2</sup>Newport Beach and Torrance stations do not report capacity values on the website, so these stations are ignored in the analysis. Three new stations were added towards the end of December 2018. They were not included in the data collection or analysis.

is normalized before cluster analysis is performed.

### 2.2. Unsupervised Temporal Clustering

Unsupervised cluster analysis has been extensively used in several domains to study categories of distinct classes occurring in the dataset, with the most common and popular algorithm being the K-Means Clustering ([Stuart, 1982](#)). While this may work well on most univariate or multivariate datasets, this density-based clustering approach has limitations when applied to time-series or spatial data as it cannot handle the dependencies across data points.

Temporal clustering is increasingly being used in finance, IoT and energy domains to identify behavior variation and anomaly detection ([Urosevic et al., 2018](#); [Hendricks et al., 2016](#)). Calculating distance measures across multiple time-series sequences is a tricky task, and can often depend on what invariance is considered to measure in the respective domains. However, Batista et al. proved that the complexity invariant distance measure is generally applicable to most time-series datasets ([Batista et al., 2011](#)). This distance measure calculates Euclidean distance between the two time-series and a correction factor is applied that factors in the differences in complexities. We use this approach to analyze the behavioral variation across the time-series fuel utilization data of hydrogen stations.

Although there is an underlying spatial component in the data, the station locations are sparse enough at this point and are considered independent of each other. We confirmed this by calculating *Moran's I coefficient* ([Anselin et al., 2006](#)) for the locations at different time intervals, and found that there is no significant spatial autocorrelation. However, as more stations are added in the future, this approach would require a spatiotemporal clustering algorithm to better understand how the location of the station contributes to their respective behavior and utilization.

### 2.3. Survey Analysis

In the third phase of the project, we designed a short, simple survey on hydrogen fueling station performance. We recorded answers from about 100 participants who currently drive FCVs in California<sup>3</sup>.

The following are the list of questions answered by the survey participants: (a) location, (b) most and least preferred stations, (c) reasons for station preferences, (d) backup stations, and (e) reasons for station avoidance. For the reasons behind the station preferences, respondents were asked to choose all the reasons that apply, such as proximity to home, proximity to work, reliability, hours of operation, safe neigh-

<sup>3</sup>Survey was hosted on <https://www.surveymonkey.com> website and the results were collected between the 5<sup>th</sup> and 15<sup>th</sup> of January, 2019.

borhood, no station congestion, ease of use, and were asked to check whichever ones apply to them. They were also encouraged to record any other comments under ‘Other (specify)').

### 3. Results

The results shown in Figure 2 indicate that there are four significant clusters based on the temporal utilization patterns of hydrogen stations. To understand what these clusters mean, we took the survey responses as well as station attributes to divide them into the following categories.

**1. Reliable stations:** The top preferred stations identified by the survey respondents as most reliable appear in cluster group **1**. These stations act as a guideline for determining whether the fueling station is performing as intended.

**2. Over-stressed stations:** The survey respondents were asked to identify the stations they found to be ‘too busy’. As cluster group **2** predominantly aligns with this answer, they are identified as stations that need intervention in terms of increase in storage capacity or a supplementary station nearby.

**3. Non-standardized Reporting:** Cluster group **3** of hydrogen stations appear to have a different storage mechanism. The liquid hydrogen fuel is pumped and stored in a bigger tank which is held at a cryogenic temperature, which is then vaporized, compressed and stored in a smaller tank for dispensing (CAFCP, 2019c). The capacity values reported are that of the smaller tank which does not reflect the actual capacity of the station. For standard comparison across stations, this reporting mechanism needs to be adjusted.

**4. Connector stations:** Two stations, Harris Ranch and Lake Tahoe are clustered together in group **4** as these are buffer stations for people traveling long distances and are not used on a regular basis.

**5. Unusual downtime:** West LA station stands out in the cluster group as there was an unusual downtime for a number of days during the time of data collection. This is not always the case during other time periods.

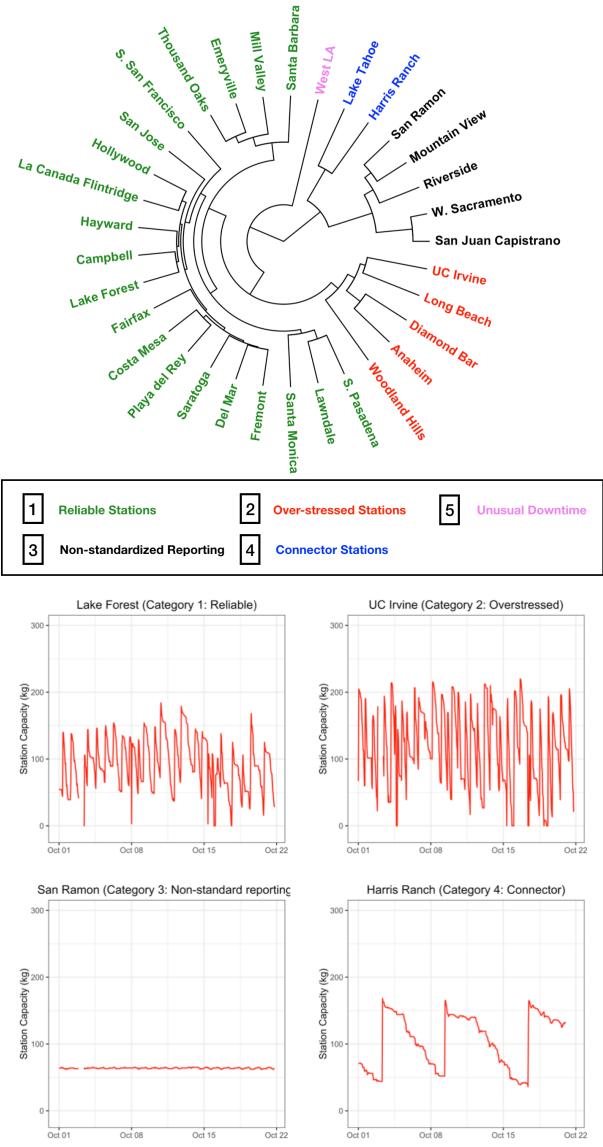


Figure 2. Visualization of the Temporal Cluster Analysis of Hydrogen Fueling Station Utilization

### 4. Conclusion

Adoption of a zero-emission vehicle fleet is a critical component to reducing carbon emissions in the transportation sector. We need diversity of ZEV technologies in the fleet to satisfy the concerns of different segments of consumers. In the case of FCVs, consumers need to trust that they would have reliable fueling infrastructure before they purchase the vehicle. Monitoring of station performance would, therefore, provide transparency to customers and help policymakers to fix issues as they arise.

We show the use of unsupervised temporal clustering algo-

rithm to determine hydrogen refueling station performance based on station capacity time-series data. The survey responses we obtained helped to understand the reasons behind our cluster categories. The analysis presented in this paper is for a static snapshot of time, hence, a machine learning algorithm is not strictly necessary as we could have obtained the same results through a heuristic. However, California plans on expanding the number of stations to about 200 stations in the year 2025, which increases the spatial and temporal complexity of the data. Therefore, a simple heuristic may fail to capture the utilization of these stations and a machine learning algorithm demonstrated in this paper may be necessary.

Even though the analysis presented in this paper focuses on a static snapshot of time, this approach could be implemented as a continuous monitoring effort with real-time streaming of station capacity data. Finally, we are releasing the hourly station capacity dataset that we collected for the machine learning researchers to further explore temporal and spatiotemporal projects.

## References

AEO. Annual energy outlook 2018: Energy-related carbon dioxide emissions by end use, 2018.

Anselin, L., Syabri, I., and Kho, Y. Geoda: An introduction to spatial data analysis. 2006.

ARB. 2018 Annual Evaluation of Fuel Cell Electric Vehicle Deployment & Hydrogen Fuel Station Network Development. Technical report, California Air Resources Board, 2018.

Batista, G. E. A. P. A., Wang, X., and Keogh, E. J. A complexity-invariant distance measure for time series. In *SDM*, 2011.

Bhatti, S. F., Lim, M. K., and Mak, H.-Y. Alternative fuel station location model with demand learning. *Annals OR*, 230:105–127, 2015.

CAFCP. By the numbers. [https://cafcp.org/by\\_the\\_numbers](https://cafcp.org/by_the_numbers), 2019a. Accessed: 2019-01-18.

CAFCP. California fuel cell partnership. <https://cafcp.org/>, 2019b. Accessed: 2019-01-18.

CAFCP. Hydrogen stations: Produce or deliver? <https://h2stationmaps.com/hydrogen-stations>, 2019c. Accessed: 2019-01-18.

Hendricks, D., Gebbie, T., and Wilcox, D. Detecting intraday financial market states using temporal clustering. *Quantitative Finance*, 16(11):1657–1678, 2016. doi: 10.1080/14697688.2016.1171378. URL <https://doi.org/10.1080/14697688.2016.1171378>.

Honma, Y. and Kurita, O. A mathematical model on the optimal number of hydrogen stations with respect to the diffusion of fuel cell vehicles. 2008.

Kang, J. E. and Recker, W. Strategic hydrogen refueling station locations with scheduling and routing considerations of individual vehicles. *Transportation Science*, 49: 767–783, 2015.

Kelley, S. Driver Use and Perceptions of Refueling Stations Near Freeways in a Developing Infrastructure for Alternative Fuel Vehicles. *Social Sciences*, November 2018.

Langley, P. Crafting papers on machine learning. In Langley, P. (ed.), *Proceedings of the 17th International Conference on Machine Learning (ICML 2000)*, pp. 1207–1216, Stanford, CA, 2000. Morgan Kaufmann.

Lim, S. and Kuby, M. J. Heuristic algorithms for siting alternative-fuel stations using the flow-refueling location model. *European Journal of Operational Research*, 204: 51–61, 2010.

Lutsey, N. Transition to a Global Zero-Emission Vehicle Fleet: A Collaborative Agenda for Governments. Technical report, The International Council on Clean Transportation, sep 2015.

Nicholas, M. A. and Ogden, J. M. Detailed Analysis of Urban Station Siting for California Hydrogen Highway Network. *Transportation Research Record*, 2006:129–139, 2007.

NREL. Hydrogen fueling infrastructure analysis. <https://www.nrel.gov/hydrogen/hydrogen-infrastructure-analysis.html>, 2019. Accessed: 2019-01-18.

Ogden, J. and Nicholas, M. Analysis of a cluster strategy for introducing hydrogen vehicles in Southern California. *Energy Policy*, 39(4):1923–1938, April 2011.

Stephens-Romero, S. D., Brown, T. M., Kang, J. E., Recker, W. W., and Samuels, G. S. Systematic planning to optimize investments in hydrogen infrastructure deployment. 2010.

Stuart, L. P. Least squares quantization in PCM. . *Information Theory, IEEE Transactions*, 28.2:129–137, 1982.

Urosevic, V., Kovacevic, , Kaddachi, F., and Vukicevic, M. Temporal Clustering for Behavior Variation and Anomaly Detection from Data Acquired Through IoT in Smart Cities. *Recent Applications in Data Clustering*, 2018.

Wahlman, A. Charger congestion a problem for electric cars: Could we already have too many electric cars?, 2013. Accessed: 2019-01-18.