A Data-Driven Methodology for Forecasting Long-Term Mission Profiles of Household Appliances from Limited Usage Data

Enrico Belmonte
Electrolux Italia S.p.A., Porcia, Italy, enrico.belmonte@electrolux.com
Martin Neumann
AB Electrolux, Stockholm, Sweden, martin.neumann@electrolux.com
Ian Marsh
Industrial Systems, RISE, Research Institutes of Sweden, AB, Sweden. E-mail: ian.marsh@ri.se

The derivation of mission profiles for household appliances is a critical step in the product development process, as accurate knowledge of consumer usage patterns enables engineers to prevent both under- and over-design. The increasing availability of connectivity in domestic appliances has created new opportunities to replace traditional assumptions, often based on surveys or expert judgment, using data-driven insights. However, understanding long-term usage behaviour remains challenging because available connectivity datasets typically cover periods much shorter than the target service life of the products. This paper presents a forecasting methodology for deriving long-term mission profiles from limited historical data, typically one year, considering univariate usage distribution. The proposed approach aims to overcome current data-length limitations and establish robust, scalable methods for predicting design-goal mission profiles in the context of connected household appliances.

Keywords: Reliability, Product Lifetimes, Prediction, Connectivity

1. Introduction

Deriving representative mission profiles is a key task in reliability engineering of home appliances. Connected devices provide detailed insight into usage behavior and patterns. However, available datasets often contain only limited early-life observations. This paper presents a distribution-based forecasting methodology to extrapolate long-term cumulative usage from short-term data. The method explicitly models seasonality, stochastic variability, and user persistency. Monthly usage is modeled with a seasonal Gamma distribution estimated by moment matching. User heterogeneity is captured through a multiplicative persistency factor.

The approach is validated by forecasting 30-month cumulative usage using only the first 12 months of data. Predictions are compared against observed ground truth. Validation is performed at the distribution level using variance decomposition and the 1-Wasserstein distance. Results show that early usage data captures the dominant source of long-term variability. This enables accurate distributional forecasts.

2. Previous work

Tecchio et. al. [1] analyse the lifetimes and failure modes of defective washing machines and dishwashers using empirical repair and warranty data . The study identifies the most common technical failures and shows that many appliances are discarded prematurely despite being repairable. Their work has broadly the same goals as ours, but we do not use warranty and repair information, rather the data from connectivity. We do not try and identify the reasons for failure as they do, design, limited repairability, and access to spare parts. Dickson et. al [2] re-examine how consumers prioritize attributes when purchasing home appliances by replicating earlier research. They show that the importance of attributes such as energy efficiency, durability, and brands varies across product categories and over time. The research results are sensitive to measurement methods and context. Their work is similar in goals, but quite different in approach; our work is data, quantitative and statistical based. Belmonte et al. [3] looked at designing reliability using data from online, connected devices. The dataset is large and spans several years, and the paper shows device usage, patterns of usage and how to deal with incomplete pictures of devices operating in the real world. This paper uses adds to Belmonte by rigorous statistical methods to estimate the lifetime of domestic goods.

3. The Dataset

Electrolux connected washing machines entered the market in 2019 and by 2025 there are hundreds of models available in over 70 countries. The connectivity data is collected in the following way. Customers connect their Electrolux appliances to the internet and register them via a smartphone app. During registration, they consent to data collection and processing according to local laws. When in use, the appliance sends real-time status updates to the IoT system enabling features like remote control and monitoring in the app. Afterward, the data are pseudonymized by removing personal information and stored in the cloud for analysis. Appliances can join or leave the service at any time and new models enter the markets constantly. As a result, the amount of data and period of time per appliance can differ widely.

For this work a sample of several thousand appliances, across all models and markets, was select that had at least 30 months of data. Data collected in this work was is inherently noisy. Appliances might suffer from bad connection; local networks might not be always active, and data could be duplicated, lost or reordered in transmission. To improve data quality appliances that showed a significant number of missing or incomplete messages where removed. While some noise remains, statistical approaches can be used to filter outliers. As a final step the data was aggregated per appliance counting the number of washing cycles for each month.

4. Statistical approach

Monthly appliance usage is observed as the number of completed cycles per appliance. Each appliance is associated with a time series of monthly counts, denoted as:

$$X_{i,m} \text{for appliance } i \text{ and month } m$$

The objective is to forecast the distribution of cumulative usage over a horizon T, defined as:

$$S_{i,T} = \sum_{m=1}^{T} X_{i,m}$$

where $S_{i,T}$represents the cumulative number of cycles performed by appliance i over the first Tmonths of usage. Forecasting is performed using only the first $H < T$months of observed data. In this study, $H = 12$ months and $T = 30$ months.

Monthly usage is modeled using a Gamma distribution:

$$X_{i,m} \sim \text{Gamma}(k_m, \theta_m \cdot Z_i)$$

where $k_m$and $\theta_m$are month-of-year–specific shape and scale parameters capturing seasonal effects, and $Z_i$is an appliance-specific persistency factor accounting for user heterogeneity. The Gamma distribution is chosen for its flexibility in modeling right-skewed count data and its analytical tractability under aggregation. Figure 1 shows three Gamma distributions modeling the number of cycles per month, corresponding to three different months of appliance operation.

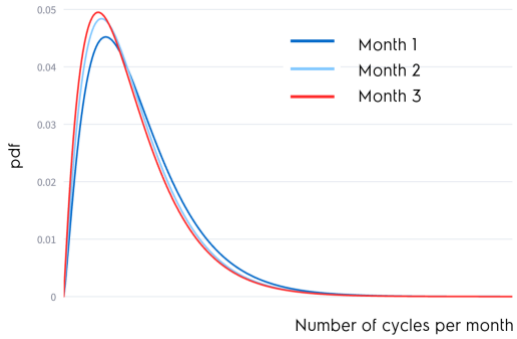

Figure 1. Estimated probability density functions of the number of cycles per month for Months 1–3 of appliance operation.

For a Gamma distribution (with loc = 0), the mean is:

$$\mu_m = k_m \cdot \theta_m$$

For each calendar month, Gamma parameters are estimated using the method of moments based on pooled observations:

$$\hat{k}_m = \frac{\mu_m^2}{\sigma_m^2}, \hat{\theta}_m = \frac{\sigma_m^2}{\mu_m}$$

where μ_m and σ_m^2 are the empirical mean and variance of monthly usage.
Moment matching is preferred over maximum likelihood estimation due to its numerical stability, robustness to discrete data, and direct interpretability in terms of mean and variance preservation.
User heterogeneity is captured through a multiplicative persistency factor Z_i, defined empirically from the first H months of usage as

$$Z_i^{(H)} = \frac{\sum_{m=1}^{H} X_{i,m}}{\sum_{m=1}^{H} \mathbb{E}[X_m]}$$

This formulation isolates user-specific intensity from seasonal effects and ensures that $\mathbb{E}[Z] \approx 1$ at the population level.
The coefficient of variation of Z, denoted as $\mathrm{CV}(Z) = \sigma_Z/\mu_Z$, is used as a compact, dimensionless measure of usage heterogeneity across the appliance population. A low $\mathrm{CV}(Z)$ indicates a relatively homogeneous user base with similar usage intensities, whereas a high $\mathrm{CV}(Z)$ reflects strong dispersion driven by the coexistence of low- and high-usage users. Appliances with $Z_i > 1$ tend to operate more frequently than the population average, while appliances with $Z_i < 1$ tend to operate less frequently. Seasonality determines when usage occurs throughout the year, whereas the persistency factor determines how intensively an appliance is used. Persistency is assumed to be stable over time, allowing user behavior observed in early life to be propagated consistently in long-term usage forecasts.

5. Forecasting Long-Term Usage Distributions

5.1 Aggregation and Variance Decomposition

The forecasted cumulative usage distribution at horizon T is derived by aggregating monthly usage contributions and applying the law of total variance:

$$\mathrm{Var}(S_T) = \mathbb{E}[\mathrm{Var}(S_T \mid Z)] + \mathrm{Var}(\mathbb{E}[S_T \mid Z])$$

where the two terms have distinct interpretations.
The first term represents stochastic month-to-month variability, arising from random fluctuations in monthly usage around the seasonal mean. The second term captures between-user heterogeneity, driven by persistent differences in usage intensity across appliances and quantified through the persistency factor Z.
This decomposition provides an important insight: as the forecast horizon increases, the contribution of monthly stochastic variability grows sublinearly and progressively averages out, whereas the contribution of user heterogeneity accumulates proportionally to the square of the horizon. Consequently, for long-term forecasts, between-user heterogeneity dominates the total variance of cumulative usage.

5.2 Forecasted Distribution Construction

Using the estimated seasonal monthly moments and the coefficient of variation of the persistency factor, $CV(Z)$, the cumulative usage distribution at horizon T is approximated by a Gamma distribution whose moments are matched analytically. The expected cumulative usage is given by

$$\mathbb{E}[S_T] = \sum_{m=1}^{T} \mu_m$$

where μ_m denotes the mean usage for calendar month m. The variance of cumulative usage is approximated as

$$\mathrm{Var}(S_T) = \sum_{m=1}^{T} \sigma_m^2 + CV(Z)^2 \cdot \mathbb{E}[S_T]^2$$

where σ_m^2 is the month-specific variance. The first term accounts for the aggregated monthly stochastic variability, while the second term reflects the amplification of variance induced by persistent user heterogeneity.
Matching these moments yields an approximation of the long-term cumulative usage distribution, without relying on individual-level trajectory predictions.

5.3 Monte Carlo Simulation

Monte Carlo simulation is used to numerically propagate uncertainty and validate the analytical approximation. The simulation requires as inputs: i) the seasonally varying Gamma parameters estimated from the training window, ii) appliance-level persistency factors capturing user heterogeneity; iii) the forecast horizon T; iv) and the number of simulation runs per appliance.
Each simulation draws possible values of monthly usage according to the model and sums them over the forecast horizon to obtain total usage values. The resulting samples

describe the forecasted usage distribution and are used to check the analytical results.

6. Validation Strategy

6.1 Data Coverage and Sampling Limitations

While the sample selection process was designed to minimize bias it had limitation due to the structure of the data. Since multiple years of data where needed for the analysis more recent appliances where not represented. While the sampling was agnostic to location and therefore, climate and cultures, its likely that the European market is over represented. It makes up a large portion of the overall population especially for the older models.

6.2 Temporal Hold-Out Design

Model validation is conducted using a temporal hold-out strategy applied at the appliance level. The training and validation datasets contain the same appliances. However, validation is restricted to appliances with at least 30 months of observed usage, so that their long-term cumulative usage can be used as ground truth. Months 1–12 are used exclusively for training, months 1–30 provide the ground-truth cumulative distribution. The appliance population is kept fixed, and training and validation are separated along the time axis. This avoids population bias and information leakage, and reflects a realistic forecasting scenario where only early-life data are available at prediction time.

6.3 Distribution-Level Validation Metrics

Forecast accuracy is evaluated at the distribution level, consistent with the objective of predicting population usage loads rather than individual appliance trajectories. Validation is based on the following criteria, applied in the following order:
i) User heterogeneity stability, assessed through the coefficient of variation of the persistency factor, $CV(Z^{(H)})$, computed at increasing observation horizons (12, 24, and 30 months);
ii) Central tendency and tail behavior, assessed through direct comparisons between forecasted and observed values of the mean and selected percentiles (P50, P90, and P95) of the cumulative usage distribution;
iii) Overall distributional agreement, quantified using the 1-Wasserstein distance between the forecasted and observed cumulative usage distributions. To enable a scale-independent interpretation, the Wasserstein distance is normalized by the mean observed cumulative usage, yielding a dimensionless measure of distributional discrepancy.

7. Results and Discussion

The empirical results provide consistent evidence in support of the proposed forecasting approach. First, the coefficient of variation of the persistency factor, $CV(Z)$, stabilizes after the first year of usage (Table 1)

Table 1. CV(Z) at different observation horizons for the same appliance population.

Months	CV(Z)
12	0.542
24	0.533
30	0.531

This indicates that user heterogeneity is observable early in the appliance lifetime and can be reliably estimated from short-term data, without requiring long observation periods.
Second, direct comparisons between forecasted and observed cumulative usage distributions at 30 months show good agreement in both central tendency and tail behavior. In particular, the forecasted mean and key percentiles (P50, P90) closely match their observed counterparts, indicating that the model accurately captures both typical and high-usage regimes (Table 2)

Table 2. Forecasted vs. Observed Cumulative Usage at 30 Months (normalized values).

	Observed S30	Forecasted S30	Relative difference
Mean	100	98	-0.26
p50	90	87	-0.43
p90	174	173	-0.12

Third, the normalized 1-Wasserstein distances between forecasted and observed 30-month cumulative usage distributions is below 5%. This confirms overall distributional agreement and provides a compact summary measure consistent with the mean and percentile comparisons. Finally, the comparison of the forecasted and observed probability density functions reveals close alignment in overall shape, including the location of the mode and the decay of the right tail. This qualitative agreement confirms that the proposed model reproduces not only summary statistics but also the full distributional structure of cumulative usage.

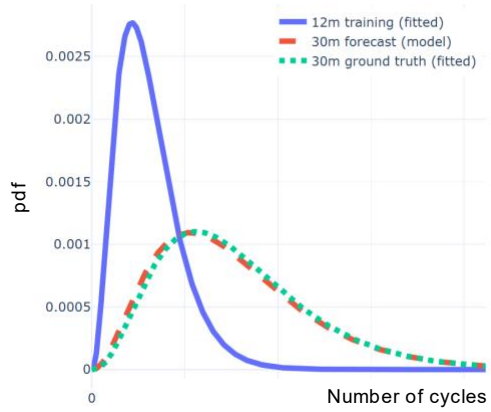

Figure 2. Probability density functions of cumulative usage after 12 months (S_{12}), observed cumulative usage after 30 months (S_{30}), and forecasted 30-month cumulative usage ($\hat{S}_{30}$)

8. Conclusions

This paper presents a methodology for forecasting long-term appliance usage distributions from short-term observations. By combining seasonal Gamma modeling, empirical persistency estimation, and variance-based validation, the approach provides reliable distributional forecasts suitable for reliability engineering.

References

1. Tecchio P., Ardente F., M. F. (2019). Understanding lifetimes and failure modes of defective washing machines and dishwashers. Journal of Cleaner Production 215, 1112–1122.
2. Dickson, P. R., R. F. Lusch, and W. L. Wilkie (1983, 03). *Consumer acquisition priorities for home appliances: A replication and re-evaluatio*n. Journal of Consumer Research 9(4), 432–435.
3. Belmonte E., Boichuk, V., Neumann M., (2025). *Leveraging usage distributions for reliability design in home appliances*. Proceedings of the 35th European Safety and Reliability & the 33rd Society for Risk Analysis Europe Conference. ESREL SRA-E 2025 Organizers. DIO: 10.3850/978-981-94-3281-3_ESREL-SRA-E2025-P0933-
4. Konishi, Sadanori, and Genshiro Kitagawa. *Information Criteria and Statistical Modeling*. New York: Springer, 2007.