Product Life Cycle Data Set: Raw and Cleaned Data of Weekly Orders for Personal Computers

Article *in* Manufacturing & Service Operations Management · May 2018

DOI: 10.1287/msom.2017.0692

5 authors, including:

Jason Acimovic
Pennsylvania State University
21 PUBLICATIONS **493** CITATIONS
SEE PROFILE

Kejia hu
Vanderbilt University
9 PUBLICATIONS **53** CITATIONS
SEE PROFILE

Douglas J Thomas
University of Virginia
16 PUBLICATIONS **109** CITATIONS
SEE PROFILE

Jan Albert Van Mieghem
Northwestern University
136 PUBLICATIONS **4,177** CITATIONS
SEE PROFILE

Product Life Cycle Data-set: Raw and Cleaned Data of Weekly Orders for Personal Computers

Jason Acimovic[1], Francisco Erize[2], Kejia Hu[3] Douglas J. Thomas[4], Jan A. Van Mieghem[5]

[1]Smeal College of Business, Penn State University

[2]Supply Chain, Dell

[3]Owen Graduate School of Management, Vanderbilt University

[4]Darden School of Business, University of Virginia

[5]Kellogg School of Management, Northwestern University

September 22, 2017

We provide and describe a data set of $N = 8935$ weekly, normalized customer orders over the entire product life cycle for 170 Dell computer products sold in North America over a three and a half year period, from 2013-2016. Total orders for these products exceeded 4 million units and well over a billion dollars in revenue. While Dell is historically known for fulfilling customer demand with a build-to-order approach, the products in this data set were designated as build-to-stock products. There are three elements in the data that, depending on the research application, researchers may want to identify or mitigate. First, some products have seemingly anomalous orders representing one-time purchases from large customers. Second, there are negative values for some products representing order cancellations. Third, end-of-life sales may be significantly influenced by management action. We present approaches for cleaning the data to address these issues. The data described in this paper are publicly available at [***please contact the first author].

1. Introduction

We provide a data set containing the normalized weekly customer order volume for 170 stock keeping units (SKUs) over their entire product life cycles (PLCs). The data set contains $N = 8935$ weekly order volumes (unique SKU-week observations) for a grand total of 4,037,826 computers sold by Dell in North America over a three and a half year period, from 2013-2016. Dell, Inc. is the third largest personal computer vendor in the world and the data set contains four product categories within personal computers: laptops, desktops, mobile workstations, and fixed workstations. These tend to be short life cycle products, with the median PLC length of the 170 SKUs equal to 51 weeks. Our primary intent in acquiring this data set was to develop methods to forecast new PLC curves (Hu et al. 2017), although the data may be useful for other purposes as well.

Dell has historically used a build-to-order approach to fulfill customer orders. Recently, to complement this approach, Dell began fulfilling demand for some selected computer configurations in a build-to-stock (BTS) manner. The intent is to select products to be fulfilled from stock where customers may value a simplified ordering process and fast delivery over the ability to customize

their product. Some BTS products are available on Dell's website under a program called Smart Selection with the stated aim to provide "a simplified ordering process for our best value, prebuilt systems custom-designed based on customer feedback."[1] Dell holds these computer configurations in its fulfillment centers, ready to ship. While this raises inventory costs, the benefits include both economies of scale in the production and transportation of higher volumes of a smaller set of SKUs and reduced lead time to the customers. The 170 products in this data set were all designated as BTS products.

We see the main contributions of this document and accompanying data set as threefold. First, this large data set of full product life cycles can be used to verify and replicate fitting and clustering results in Hu et al. (2017), a desirable step in the scientific process. Second, the data set may be useful to the research community for studying various questions relating to the product life cycle. Some ideas are proffered in Section 4 below. Third, we also view the documentation of the cleaning and preparation of the data as a contribution. Often, data from the real world is messy for a variety of reasons. By documenting how we cleaned and prepared this data (which was informed and advised by Dell demand planners), data analysts and researchers can utilize similar techniques where appropriate in analyzing their own data sets.

2. Description of data

Dell demand planners provided us with data sets of customer orders for 170 BTS products between April 2013 and October 2016. A product or SKU corresponds to a built computer with a completely specified configuration such as a Dell Optiplex 'i5/8GB/500GB/Gfx/RW' where the configuration 'x/y/z/a/b' specifies the 'Intel processor type/amount of RAM/capacity of hard drive/Graphics Processor/Optical read-write drive.'[2] For each of the SKUs, for each calendar week within its entire product life cycle, we know how many net customer orders were placed. The data file, `PLCweeklyorders_2017-09-19.csv` is available at [***please contact the first author]. Descriptions of the columns are in Table 2 below as well as a file, `PLCDataSet_DataDictionary_ColNamesAndDescriptions.txt`, also available at [***please contact the first author].

The data we have is the same data available to Dell's demand planners during the given time window. We note here some limitations.

[1] `http://www.dell.com/learn/us/en/04/campaigns/smart-select-consumer?c=us&l=en&s=dhs` accessed on September 15, 2017.

[2] This particular configuration is a typical example of a BTS product. As of September 15, 2017, this was an 'in stock ready-to-ship' offer with a price of $789. `http://www.dell.com/en-us/work/shop/desktops-workstations/optiplex-5050-tower-small-form-factor/spd/optiplex-5050-desktop`

1. Net customer orders only: For each week, the data set contains the sum of total orders placed minus returns and cancellations in that week. We do not know the true total customer orders in a given week, nor do we know the breakdown (whether it was one large order or several small orders). Additionally, if a large order is placed in one week and returned or cancelled the next week, it may lead to negative net customer orders observed in the following week. We discuss how we treat this below in Section 3.

2. Censored demand: The data set contains only customer orders and we do not have access to inventory information. From the customer point of view, if an item is not in a fulfillment center she will not see the item as explicitly out of stock. Rather the quoted lead time will be longer for items which are not in the fulfillment center. The customer may or may not decide to continue placing her order.

In the accompanying data set, we disguise the true order volumes. We normalize each SKU's lifetime demand to equal 1. As we note below, the raw data may include negative numbers due to cancellations. In these cases, the lifetime volume is still normalized to 1, so that the absolute value of each week's orders may be greater than 1. Although we disguise absolute volume information, we still provide the quintile of each SKU's average weekly volume relative to the set of 170. Average weekly volume is the measure used by Dell to define "fast-" or "slow-moving" products. Quintile 1 denotes the 34 SKUs with the highest average weekly order volumes while Quintile 5 denotes the 34 SKUs with the lowest average weekly order volume. Table 1 shows summary statistics for the entire data set before cleaning and preparation. In order to disguise total lifetime volumes further, we round each quintile's normalized weekly order volume as such:

$$RoundedNumber = \frac{[FullPrecisionNumber \times \alpha_q]}{\alpha_q},$$

where q denotes the quintile, $[\cdot]$ denotes rounding to the nearest integer, and $\alpha_q = (50000, 20000, 10000, 5000, 2000)$ for Quintiles 1 through 5 respectively[3]. This small perturbation protects sensitive volume information while preserving the integrity of the PLCs.

Below, in Section 3, we describe how we prepare and clean the data. Specifically, the steps are: detecting cancellations, adjusting for seasonality, excluding build-to-order orders, and end-of-life truncation. We report the resulting customer order data at each one of these steps. After the last step (end-of-life truncation) is performed, we re-normalize the data to sum to 1 for the product's life cycle. Table 2 provides the names and descriptions of the columns of the accompanying data set, while Figure 1 shows sample data for a hypothetical SKU. The columns in Figure 1 map one-to-one with the columns we provide in the accompanying data set, with the exception of the *TrueVolume* field which is provided here for illustrative purposes only.

[3] Each α_q is the largest number less than $MedianWeeklyOrders_q \times MedianPLCWeeks_q$ (found in Table 1) whose first digit is a factor of 10 and whose remaining digits are zeroes. Data cleaning is performed on the raw values, and rounded afterwards.

Quintile	PLC length in weeks			Weekly net customer orders
	1st Quartile	2nd Quartile	3rd Quartile	
1	48	53	85	1,021
2	47	58	75	456
3	34	49	62	267
4	23	54	61	166
5	22	39	58	66

Table 1 Summary statistics of the PLC length and order volume data before preparation, at the quintile level. (The 'weekly' net customer orders are the median of the mean weekly volumes of the SKUs in each quintile.)

Column Name	Description
Week	*The undisguised calendar date corresponding to the order week.*
SKU_ID	*A disguised identifier corresponding to each unique SKU.*
Quintile	*The quintile of the order volume for each of the 170 SKUs. Quintile 1 denotes the 34 SKUs with the highest average weekly order volumes while Quintile 5 denotes the 34 SKUs with the lowest average weekly order volume.*
RawVolume	*Normalized order volume. For each SKU, the weekly order volumes divided by the lifetime order volumes. The sum of each SKU's* RawVolume *across weeks should equal 1.*
DetectCancellations	RawVolume *column with negative order cancellations adjusted, as described in the text.*
RemoveBuildToOrder	DetectCancellations *column with large build-to-order customer orders adjusted, as described in the text.*
TruncateEndOfLife	RemoveBuildToOrder *column with the end-of-life orders truncated, as described in the text.*

Table 2 Dictionary of data file column names and descriptions.

3. Cleaning and preparing the data

We outline the data preparation steps below. The insights and practices of a demand planner at Dell guided the identification of the root cause and the proper treatment of these phenomena.

Detecting cancellations. In the raw data, we observe 62 negative net customer orders for various products out of 8935 SKU-week observations. Let D_t^i be the raw observed net customer order value in week t for product i. T_i is the length of the life cycle (in weeks) of product i, and $t \in \{1, \ldots, T_i\}$ is the product-specific week relative to that product's launch week. The total observed customer orders for a product in week t is the sum of all the actual customer orders minus all the order cancellations and returns for that week. Returns are relatively rare, and the negative

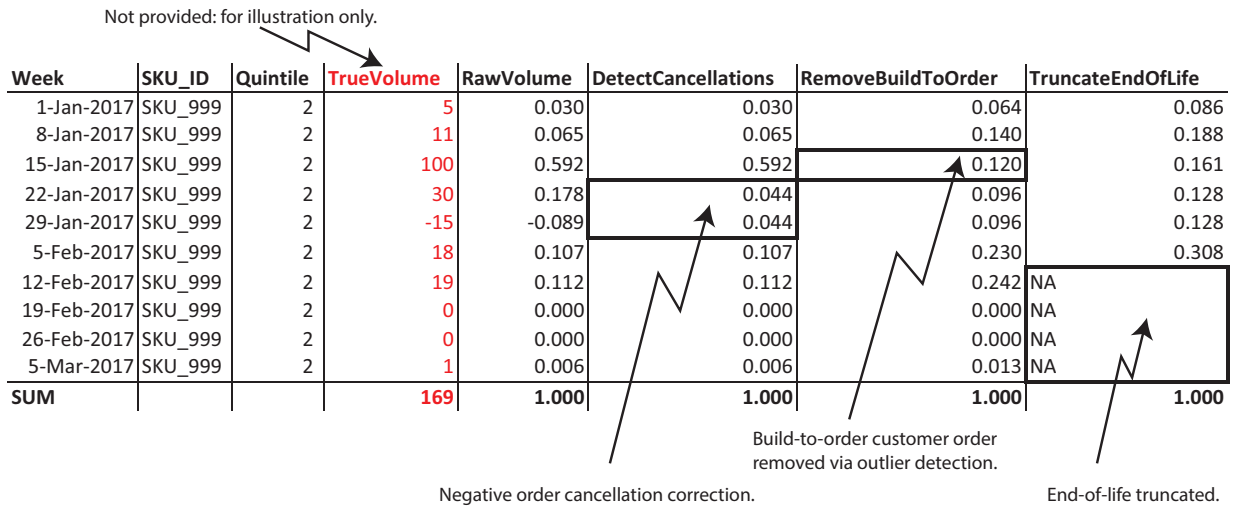

Figure 1 **Illustration of the data we provide for a hypothetical SKU. The numbers are consistent, in that applying our data preparation steps on the hypothetical SKU's normalized raw order data will produce the data shown in the rightmost three columns. The $TrueVolume$ column and the SUM row are not provided in the data set.**

data points suggest that an order which was placed in an earlier week is being cancelled. We 'correct' these cancellations by implementing the following algorithm which smooths out negative orders while keeping constant the total lifetime order amount. First, we move from the last period T to the first period 1 time-step by time-step. If we come across a negative order at time t, then we set this time period's order and the previous period's order equal to the simple average of their old orders: $D_t^{Pass[1],i} = D_{t-1}^{Pass[1],i} = \frac{D_t^i + D_{t-1}^i}{2}$, where $Pass[1]$ denotes we are moving backwards in time on the first iteration. After this first pass, negative orders may still exist: for instance, if a large negative cancellation was really tied to an order placed more than one week ago. If this is the case, then move forward in time time-step by time-step. If a negative order is encountered on this pass at time t, set $D_t^{Pass[2],i} = D_{t-1}^{Pass[2],i} = \frac{D_t^{Pass[1],i} + D_{t-1}^{Pass[1],i}}{2}$, where $Pass[2]$ indicates this is the second pass overall, and the first forward pass. If necessary, repeat m times until no negative orders are found. Then, set $D_t^{C[1],i} = D_t^{Pass[m],i}$, where $C[1]$ denotes cleaning step number 1. Figure 2 shows four SKUs on which we apply negative order treatment.

Adjusting for seasonality and normalization. Products in our data set were launched at various times throughout the calendar year, and this information is summarized in Table 3. Fewer than 3% of the SKUs in our dataset have product life cycles greater than two complete years, so we cannot estimate any seasonal effects at the product level for most of our products. Since there is variation in the launch timing, and the data set covers more than two years, it may be possible to disentangle seasonal effects from PLC behavior at an aggregate level. However, products in different

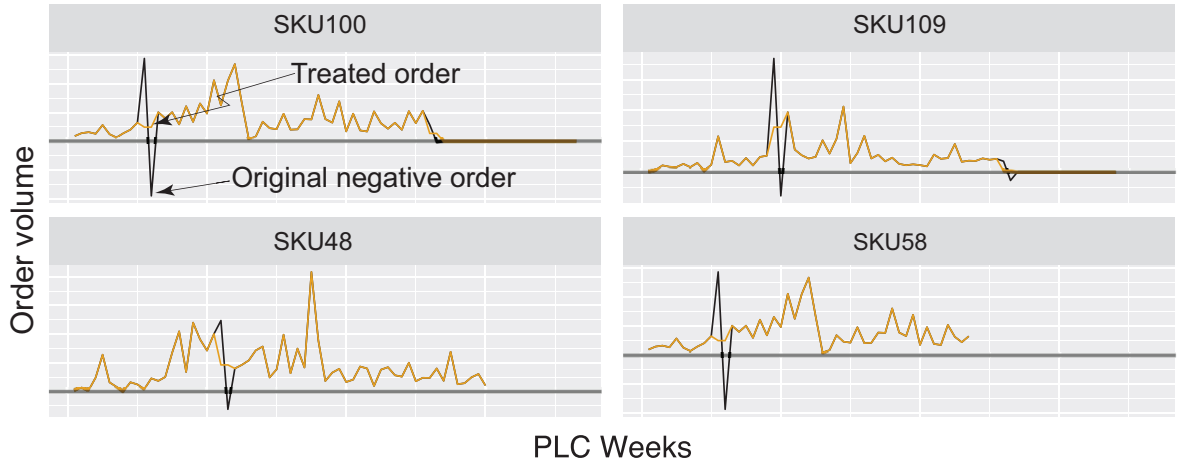

Figure 2 **Illustration of our negative order correction method which essentially smooths a negative outlier with a previous large order.**

Month	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct	Nov	Dec
Number launched	13	44	3	2	2	26	21	3	17	11	24	4

Table 3 **Distribution of number of new products' launches across different calendar months. Launches are distributed across different months.**

categories for different markets may have different seasonal patterns. For instance, different industrial customers and governmental organizations have different fiscal years; thus the phenomenon of 'end-of-fiscal-year-buying' will occur in different (but perhaps fixed) months throughout the year. Since we do not have information about the intended market for each product we do not attempt to estimate seasonal factors for subsets of products. For these reasons, properly adjusting for seasonality in this data set is challenging, and we do not include deseasonalized data with this data set. Hu et al. (2017) adjust for seasonality at the aggregate level using a monthly, multiplicative model and find that adjusting for seasonality hurts out of sample forecast performance.

Excluding build-to-order customer orders (Outlier correction). Very large orders are quite likely to be one-time orders placed by a large customer, and it is probable such orders would have been fulfilled in a build-to-order manner. If a researcher is interested in the orders that were fulfilled from stock, it may be beneficial to identify and remove these large orders. We do not know exactly which portion of a week's customer orders was due to these very large orders. As a proxy, we identify weeks with very large customer order totals using outlier detection, assuming that this outlier is actually mostly made up of a very large order with a different workstream. We replace these outliers with 'reasonable' values (defined below) because once the build-to-order workstream units are removed, there are still likely underlying BTS customer orders. We identify outliers by

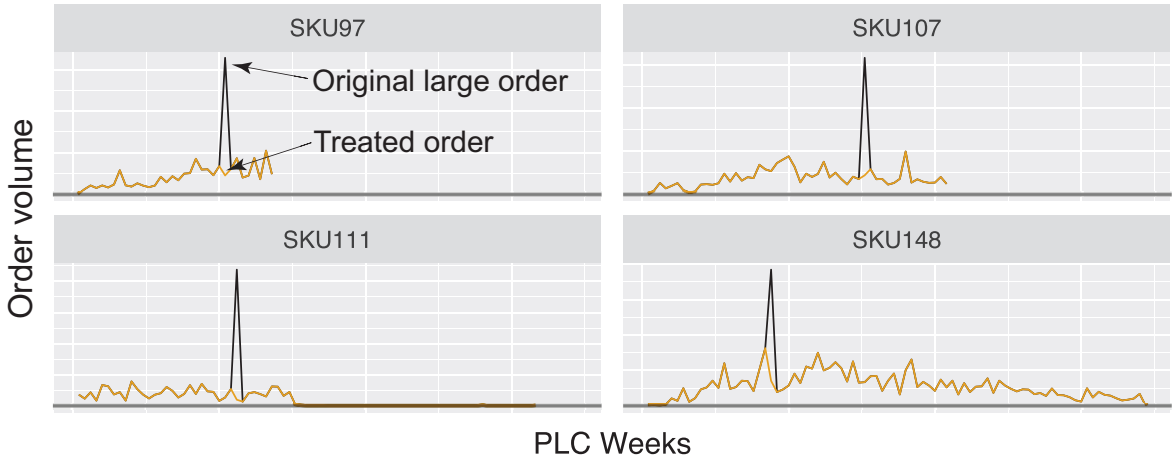

Figure 3 **Illustration of our large-order treatment method on four SKUs. The outliers denoted by the black lines (one outlier in each panel) are detected and replaced by the value as defined by Equation (1).**

the time series outlier detection method described in Chen and Liu (1993)[4]. In essence, the method outlined in that paper fits a time series model to the data and then identifies outliers significantly deviating from this time series model. Thirteen out of 8935 orders are identified as outliers across 13 SKUs. We replace the detected outliers for product i at week t with their weighted moving averages, as outlined in Roberts (2000). Figure 3 shows examples of outlier correction for four SKUs. Recalling that T_i denotes PLC length of product i, we set:

$$D_t^{C[2],i} = \frac{D_1^{C[1],i} + 2D_2^{C[1],i} + \cdots + (t-1)D_{t-1}^{C[1],i} + (T_i - t)D_{t+1}^{C[1],i} + \cdots + D_{T_i}^{C[1],i}}{(1 + \cdots + t - 1) + (1 + \cdots + (T_i - t))}. \tag{1}$$

End-of-life truncation. Customer orders near the end of the life cycle can be strongly influenced by managerial decisions such as promotions, clearance sales, or timing of the introduction of a new product intended to replace an old one. Additionally, in our data set, the length of the product life cycle can be artificially extended past when a product's life has essentially ended, since a single customer order may occur or a return or cancellation may be made weeks later. Figure 4 shows the first behavior (managed end-of-life) in the left panel and the second behavior (artificially extended PLC) in the right panel.

If a researcher wishes to focus on "naturally occuring" demand during the product life cycle, rather than orders that occur during an actively managed or artificially extended end-of-life, one should exclude customer orders near the end of the life cycle.

[4] We utilize the methods of Chen and Liu (1993) specifically to detect Additive Outliers, and set the critical value C equal to 10.0.

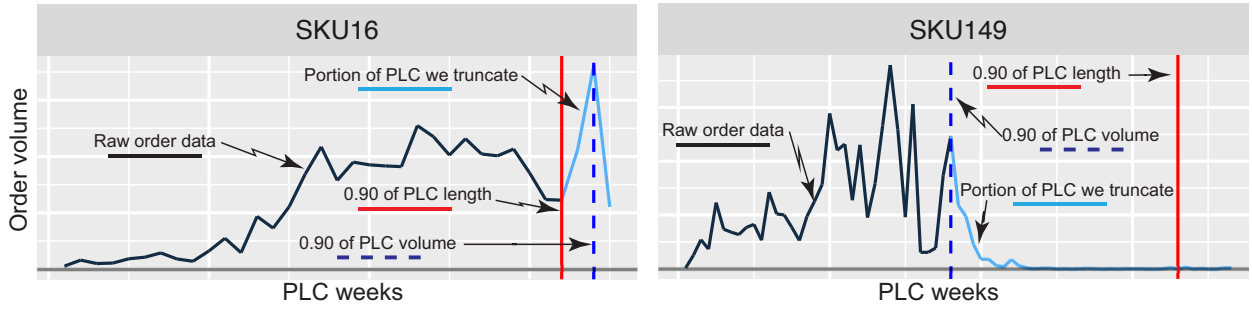

Figure 4 **Illustration of our end-of-life truncation approach. The solid red line represents the cut-off based on PLC length, while the dashed blue line represents the cut-off based on volume. Our method cuts off all data points that occurred after the earliest cut-off point from either method.**

Let θ^t be the fractional portion of the length (in weeks) of the beginning of the life cycle we keep, and let θ^v be the fractional portion of the beginning life cycle order volume (in units) we keep. Specifically, for the time or volume cutoff, we include as many weeks as possible without exceeding θ^t or θ^v respectively. We initially set $\theta^t = \theta^v = 0.9$ for this data set. Figure 4 shows how the time cutoff and the volume cutoff affect the raw data, where the time cutoff dominates in the left panel and the volume cutoff dominates in the right panel. The processed data is now defined as $D_{t'}^{C[3],i} = D_t^{C[2],i}$ for t' satisfying the above criteria. T_i is redefined to be a smaller value as appropriate to account for the cut off data points.

A note on chronological ordering of data processing steps. We perform the processing steps above such that we first remove the most spurious data and step-by-step work with cleaner and cleaner data. First, the negative order detection corrects for customer requests that were never (or erroneously) placed. Second, we suggest performing de-seasonalization before excluding build-to-order customers because strong seasonality may make typical demand appear as an outlier to the outlier-detection algorithm (or may even make a true outlier appear as typical demand). We perform end-of-life truncation last as this is less a 'cleaning' step (as every data point represents actual de-seasonalized customer orders that occurred) and more of a step to make the data usable.

4. Discussion

We provide and describe a data set of weekly orders for computers over their entire product life cycle. The data set may be useful to the research community for studying various questions relating to the product life cycle. Some anticipated questions include: (1) How to model the PLC? Researchers have used data sets of product life cycles from other industries such as grocery (Headen 1966), food (Buzzell and Nourse 1967) and chemicals (Frederixon 1969) to empirically investigate how different functional forms fit PLCs. We are not aware of another publicly available data set that would enable further empirical research in modeling the PLC. (2) How do forecasting algorithms

perform over this data set? As noted above, Hu et al. (2017) use this data set to investigate approaches for forecasting. Researchers can now verify and replicate their fitting and clustering results, a desirable step in the scientific process. Ban et al. (2017) develop an approach to model and manage procurement over the PLC using historical PLC data. (3) What is empirically observed end-of-life behavior? (4) How to manage inventory over the PLC? Kurawarwala and Matsuo (1996) use computer sales, from a make-to-order setting, to jointly fit the PLC and make procurement decisions.

The data set, which has actual calendar dates, can also be complemented with external data sets or company information such as release dates of Intel/AMD chipsets, consumer confidence, or competitor product or stock data. This may allow the study of product launch timing, cannibalization, and market share.

This data set has several desirable attributes for researchers seeking to study issues related to order behavior over the PLC. First, this is a large data set including more than 4 million weekly unit orders over three and a half years representing over 1 billion dollars in revenue. Next, as summarized in Tables 1 and 3 there is diversity in the products in the data set in terms of the weekly volume, length of the PLC, and launch month. In addition to describing the raw data, we describe steps for cleaning the data. The cleaned data has been used to study forecasting over the product life cycle (Hu et al. 2017). While such cleaning may be desirable for certain research settings, for others it may not. Therefore, we provide the normalized data before and after each potential cleaning step.

Making this data publicly available does come with limitations due to confidentiality restrictions by our partner company: we cannot provide unscaled volumes, product attributes, or product segments. In addition, we do not have item-specific, let alone temporal, price data. However, we were able to provide volume and PLC lengths by quintile so that the normalization protects confidentiality without giving up "too much." For example, one can separately analyze aspects of the PLC for high and low volume products, potentially identifying similarities or differences across those groups.

We hope this data set will be useful to researchers seeking to empirically study aspects of the PLC. Furthermore, we hope that our approach of making a data set publicly available to the research community while still protecting company confidentiality will inspire other researchers to seek a similar "win-win" publication agreement.

References

Ban, Gah-Yi, Jérémie Gallien, Adam J. Mersereau. 2017. Dynamic procurement of new products with covariate information: The residual tree method. SSRN Scholarly Paper ID 2926028, Social Science Research Network, Rochester, NY. URL `https://papers.ssrn.com/abstract=2926028`.

Buzzell, Robert Dow, Robert E Nourse. 1967. *Product innovation in food processing, 1954-1964*. Harvard University Press, New York.

Chen, Chung, Lon-Mu Liu. 1993. Joint estimation of model parameters and outlier effects in time series. *Journal of the American Statistical Association* **88**(421) 284–297.

Frederixon, Martin Shelton. 1969. An investigation of the product life cycle concept and its application to new product proposal evaluation within the chemical industry. Ph.D. thesis, Michigan State University.

Headen, Robert Speir. 1966. *The Introductory Phases of the Life Cycle for New Grocery Products: Consumer Acceptance and Competitive Behavior*. Graduate School of Business Administration, George F. Baker Foundation, Harvard University.

Hu, Kejia, Jason Acimovic, Francisco Erize, Douglas J. Thomas, Jan A. Van Mieghem. 2017. Forecasting new product life cycle curves: Practical approach and empirical analysis. *forthcoming, Manufacturing & Service Operations Management* .

Kurawarwala, Abbas A, Hirofumi Matsuo. 1996. Forecasting and inventory management of short life-cycle products. *Operations Research* **44**(1) 131–150.

Roberts, S.W. 2000. Control chart tests based on geometric moving averages. *Technometrics* **42**(1) 97–101.