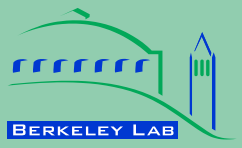
LBNL-6989E



ERNEST ORLANDO LAWRENCE
BERKELEY NATIONAL LABORATORY

Estimating Sales and Sales Market Share from Sales Rank Data for Consumer Appliances

Samir Touzani and Robert Van Buskirk

Energy Technologies Area
Lawrence Berkeley National Laboratory
Berkeley, CA 94720

February 2015

This work was supported by the U.S. Department of Energy under Lawrence Berkeley National Laboratory Contract No. DE-AC02-05CH11231.

Disclaimer

Estimating Sales and Sales Market Share from Sales Rank Data for Consumer Appliances

Abstract

Our motivation in this work is to find an adequate probability distribution to fit sales volumes of different appliances. This distribution allows for the translation of sales rank into sales volume. This paper shows that the log-normal distribution and specifically the truncated version are well suited for this purpose. We demonstrate that using sales proxies derived from a calibrated truncated log-normal distribution function can be used to produce realistic estimates of market average product prices, and product attributes. We show that the market averages calculated with the sales proxies derived from the calibrated, truncated log-normal distribution provide better market average estimates than sales proxies estimated with simpler distribution functions.

Table of Contents

1. Introduction 5
2. From Sales Ranks to Sales Volume 6
3. Data 8
4. Empirical Results of Fitting the Distribution Functions 9
5. Forecasting Sales 14
6. Application to Calculation of Market Average Quantities 16
7. Application to Calculation of Price Indices 18
8. Conclusion 20

1. Introduction

Internet-based data collection is beginning to revolutionize market and economics research. One area of activity is called "Big Data" analysis and consists of collecting very large amounts of data from the Internet and exploring a large number of inferential relationships that may exist in the data (Mayer-Schönberger and Cukier 2013). Another related, but more specialized form of Internet-based data collection is called "Scraped Data." Scraped data are available from on-line websites and databases and are collected by means of specialized computer programs that either "crawl" the websites and parse the web page text (Cavallo 2012) or interface more directly to back-end databases through the utilization of an application programmer interface (API) made available by the website providers (Chang et al. 2006).

One very important category of scraped data is product prices and attributes, which are available from on-line retailers. These prices, which are available in real-time, are beginning to find application to real-time price monitoring, especially for the calculation of price indices (Cavallo 2013). With the increasing availability and use of such product market and sales data collected over the Internet, it becomes important to estimate the sales volume corresponding to different product models and product offers observed on Internet websites. Sales volumes estimates are necessary for assigning appropriate weights to different sales offers when product market average quantities are calculated.

The price index calculations by Cavallo (2012) based on scraped data use weights when aggregating price indices from different product categories, but do not use quantity weights for individual products within a product category. However, several Internet sources provide sales ranks for individual products relative to other products in their category. Sales rank data can be included in the scraped data acquired from the Internet when prices and product attributes are collected. Conceptually these sales rank data can be used to develop quantity proxies that can be incorporated into price index calculations. When sales rank can be used to estimate sales volume this can provide a means of estimating weights for the different model-specific price changes which enable the calculation of price indices that utilize these weights. This approach has been used to measure book prices, but has not yet been generalized to other products (Chevalier and Goolsbee 2003).

In this study, we investigate the details of how sales quantity estimates can be constructed from sales rank data for household appliances. This study aims to improve the derivation of quantity proxies for market data acquired over the Internet by refining the modeling of the relationship between sales rank and sales volume. Specifically, we use detailed point-of-sales (POS) data for refrigerators, freezers, and clothes washers to estimate improved distribution functions for model-by-model sales.

A review of the literature shows that both power law functions and log normal distribution functions have been used to approximate economic distributions of sales quantity (Pinto et al. 2012 and Newman 2005). Chevalier and Goolsbee (2003) used a power law function to describe online books sales, while Hisano and Mizuno (2011) observed that the sales distribution of consumer electronics follows both power law and log-normal distributions. Stanley et al. (1995) used a log-normal distribution to fit the size distribution of firms.

In this work we empirically show that the log-normal distribution (specifically the truncated version) produces an accurate approximation of sales using sales rank data in the context of appliances. We then demonstrate the application of this improved distribution function to providing quantity proxies for the estimation of market average product prices and product attributes (specifically appliance capacity). We also examined power law distributions, but their performance was consistently inferior to the simple log normal distribution, and we do not report the details here.

When there are no limitations to data access, the data-weighting technique commonly used is to weight the data for each model in proportion to the actual sales of that model. If there are no data for estimating quantity proxies, then the simplest alternative data-weighting method is to calculate market averages weighting the values for each product model equally. We show that the method of giving each model an equal weight performs poorly in estimating market average prices or market average appliance attributes or prices compared to estimates of sales-weighted market averages.

Finally we show that the parameters of the improved distribution function are sufficiently stable, that a distribution function calibrated with historical POS data can be used to forecast sales weights from future sales rank data. Specifically, we show that, if one calibrates the parameters of the truncated log normal distribution function with historical POS data, then it is possible to accurately estimate market average quantities (such as market average price or market average appliance capacity) with future sales rank data only. The market average estimates are made by using the truncated log normal distribution function to estimate the sales quantity proxy from the sales rank data. This method of estimating market averages from sales rank can provide an accurate approximation of actual sales-weighted market average quantities if sales ranks are known with accuracy and if the distribution function parameters can be estimated.

The paper is organized into the following sections. Section 2 introduces our method to compute sales proxies using the sales rank. Section 3 provides an overview of the data used in this study. Section 4 presents the empirical results of fitting the distribution function to POS data. Section 5 compares the performance of a log normal distribution with the performance of a truncated log normal distribution for forecasting sales from sales rank data and shows that a truncated log normal distribution provides improved performance. In Section 6 we demonstrate the application of the distribution function to the estimation of market average quantities from sales rank data, and demonstrate the superior performance of the truncated log normal distribution function. In Section 7 we demonstrate the application of our method for estimating sales proxies to the calculation of a weighted version of an online price index.

2. From Sales Ranks to Sales Volume

If we consider sales quantity as a random variable S and let $(s_1,...,s_N)$ be a set of N observations, then if the data are observed in rank order, i.e. the largest is first and the smallest is last, then the function that provides the observed rank to sales can be written in the following form:

$$F_S(s_i) = 1 - \frac{r_i}{N} \qquad (1.1)$$

where s_i is the quantity of sales of the *i*th model, r_i the corresponding sales rank, N the number of considered appliance models, and $F_X(s_i)$ corresponds to the cumulative distribution function defined as

$$F_S(s) = P(S \le s) = \int_{-\infty}^{t=s} f(t)\,dt \qquad (1.2)$$

where *f(t)* is the probability density function. The formulation (1.1) transforms the sales rank into the cumulative distribution of the sales quantities, and it means that there are r_i models, which have the sales quantities equal or higher than s_i .

As we will empirically show in Section 4 the log-normal distribution (LN) provides a good agreement with the sales distribution for the studied appliances. Thus, for the log-normal distribution the relation (1.1) is equivalent to

$$F_S(s_i) = \Phi\left(\frac{\ln s_i - \mu}{\sigma}\right) = 1 - \frac{r_i}{N} \qquad (1.3)$$

where Φ is the cumulative distribution function of the standard normal distribution, while the parameters μ and σ are respectively the mean and the standard deviation of the normally distributed random variable *ln S*.

It is straightforward to derive from (1.3) the estimate of the sales quantities

$$\hat{s}_i^{LN} = \exp\left(\sigma\ \Phi^{-1}\left[1 - \frac{r_i}{N}\right] + \mu\right) \qquad (1.4)$$

To compute $\hat{s}_i^{LN}$ we need to estimate the parameters σ and μ. The method of choice to estimate these parameters is the maximum likelihood estimation (MLE), which provides a unified method to estimate parameters in a parametric model (Johnson and Balakrishnan 1994 and Wasserman 2004).

One drawback of using the LN distribution to approximate the sales is the fact that a LN distribution is defined over positive real numbers, whereas sales are non-negative integers. To bypass this problem, we discretize the distribution results by mapping $\hat{s}_i^{LN}$ to the largest previous integer using the floor function, as follows:

$$\hat{s}_i^{LN} = \begin{cases} \lfloor \hat{s}_i^{LN} \rfloor, & if \quad \hat{s}_i^{LN} \ge 1 \\ 1, & else \end{cases} \qquad (1.5)$$

where $\lfloor x \rfloor = \max\{m \in \mathbb{Z} \mid m \le x\}$. Note that the second condition in (1.5) corresponds to the smallest sales quantity being one unit per time interval. When any model available in the market has no unit sales, the appliance model drops out of the distribution sample.

As we will empirically show in Section 4 the sales volume of the higher ranked models are poorly approximated by the LN distribution, which usually over-estimates the sales of the top-ranked model. In addition the smallest value of the sales volume is equal to 1. Because of these we empirically demonstrate that a truncated version of the LN distribution produces a better fit. A

truncated log normal limits the maximum and minimum sales volume in the distribution function. Thus, if we consider that *S* has a doubly truncated LN distribution, (1.1) will be equivalent to

$$\frac{F_S(s_i)-F_S(a)}{F_S(b)-F_S(a)}=1-\frac{r_i}{N} \qquad (1.6)$$

where $F_S(a)$ and $F_S(a)$ are respectively the value of the log-normal cumulative distribution function of the lower truncation bound *a* and the upper truncation bound *b*.
Thus the estimate of the sales quantities using the truncated TLN distribution is given by

$$\hat{s}_i^{TLN}=\exp\left(\sigma\ \Phi^{-1}\left[\alpha\left(1-\frac{r_i}{N}\right)+\beta\right]+\mu\right) \qquad (1.7)$$

with

$$\alpha=\Phi\left(\frac{\ln b-\mu}{\sigma}\right)-\Phi\left(\frac{\ln a-\mu}{\sigma}\right) \qquad (1.8)$$

and

$$\beta=\Phi\left(\frac{\ln a-\mu}{\sigma}\right) \qquad (1.9)$$

We set here the upper truncation bound equal to the sales quantity of the appliance model of rank 1 and the lower bound equal to 0.5; this value provides the best fitting results. The parameters σ and μ are estimated by maximum likelihood estimation (MLE). As previously done for the log-normal estimate we map the sales quantities to the largest previous integer.

$$\hat{s}_i^{TLN}=\begin{cases}\lfloor \hat{s}_i^{TLN} \rfloor, & if \quad \hat{s}_i^{TLN}\geq 1\\ 1, & else\end{cases} \qquad (1.10)$$

These equations provide the formulation of the truncated log-normal distribution that we use to model the relationship between sales rank and sales volume.

3. Data

In this work we use POS data from the NPD Group for three products: refrigerators, freezers, and clothes washers. These data are collected from a sample of U.S. retailers; for each model the data include the monthly total revenue and the unit sales volume. (A list of participating retailers can be found in Spurlock (2014), and these include some online retailers). The data also contain product characteristics for some models. The covered periods are from January 2007 to November 2011 for refrigerators and freezers and from January 2004 to December 2009 for clothes washers. The total numbers of observations are approximately 163,000 for the refrigerators, 31,000 for the clothes washers, and 12,000 for the freezers. However, for each product a significant subset of model numbers are not specified in order to ensure the anonymity of the retailers. Thus, for this reason, approximately 10% of the refrigerators and freezers observations and 15% of clothes washers observations are omitted.

For each product we divided the data into two periods: a learning period and a forecasting period, with the forecasting period representing the last 12 months of the available data. We use the learning period to estimate the distributions parameters μ^{LN}, σ^{LN}, μ^{TLN} and σ^{TLN}. The forecasting period is used to demonstrate the potential accuracy of the method of using sales rank and our estimated distribution function to approximate sales. For the learning period, the approximate mean values (and the standard deviations) of the number of models of refrigerators, clothes washers, and freezers per time period are respectively: 2,704 (993), 176 (32), and 183 (59). For the forecasting periods these statistics are respectively: 2,130 (72), 218 (8), and 162 (8). We note that there is a higher variability of the number of models during the learning periods.

Below, we provide the price averages and the capacity averages of the considered products during the forecasting period. However, for some models the NPD data do not provide the capacity and monthly revenue, and so must be dropped from the observations that were used. Thus, an additional 1% of the refrigerator observations, 12% of the clothes washer observations, and 15% of the freezer observations are omitted from the analysis described below.

4. Empirical Results of Fitting the Distribution Functions

In this section we investigate the performance of using the log-normal and truncated log-normal distributions to estimate the sales volumes of refrigerators, freezers, and clothes washers. To compute the maximum likelihood estimates of the distributions parameters, we used the R package "fitdistrplus" (Delignette et al. 2014). The distribution function is fit to the quantity of sales data for each time step, and the statistic R^2 is measured for the log data as a metric of the goodness of fit. The formula for R^2 is given by:

$$R^2 = 1 - \frac{1/N \sum_{i=1}^{N} \left(\ln s_i - \ln \hat{s}_i\right)^2}{Var\left(\ln s_i\right)} \qquad (1.11)$$

where *Var* corresponds to the variance, s_i and $\hat{s}_i$ are respectively the actual sales and model function sales estimate for rank r_i.

We illustrate our results with a Zipf plot to illustrate both the data and the fitted distribution functions. The Zipf plot is a double logarithmic scale graphical representation of the rank versus the variable, which in our case is sales. In this evaluation we use the data corresponding to the last available time step for each product. There are 2,025 refrigerator models and 153 freezer models in November 2011, and 153 clothes washer models in December 2009.

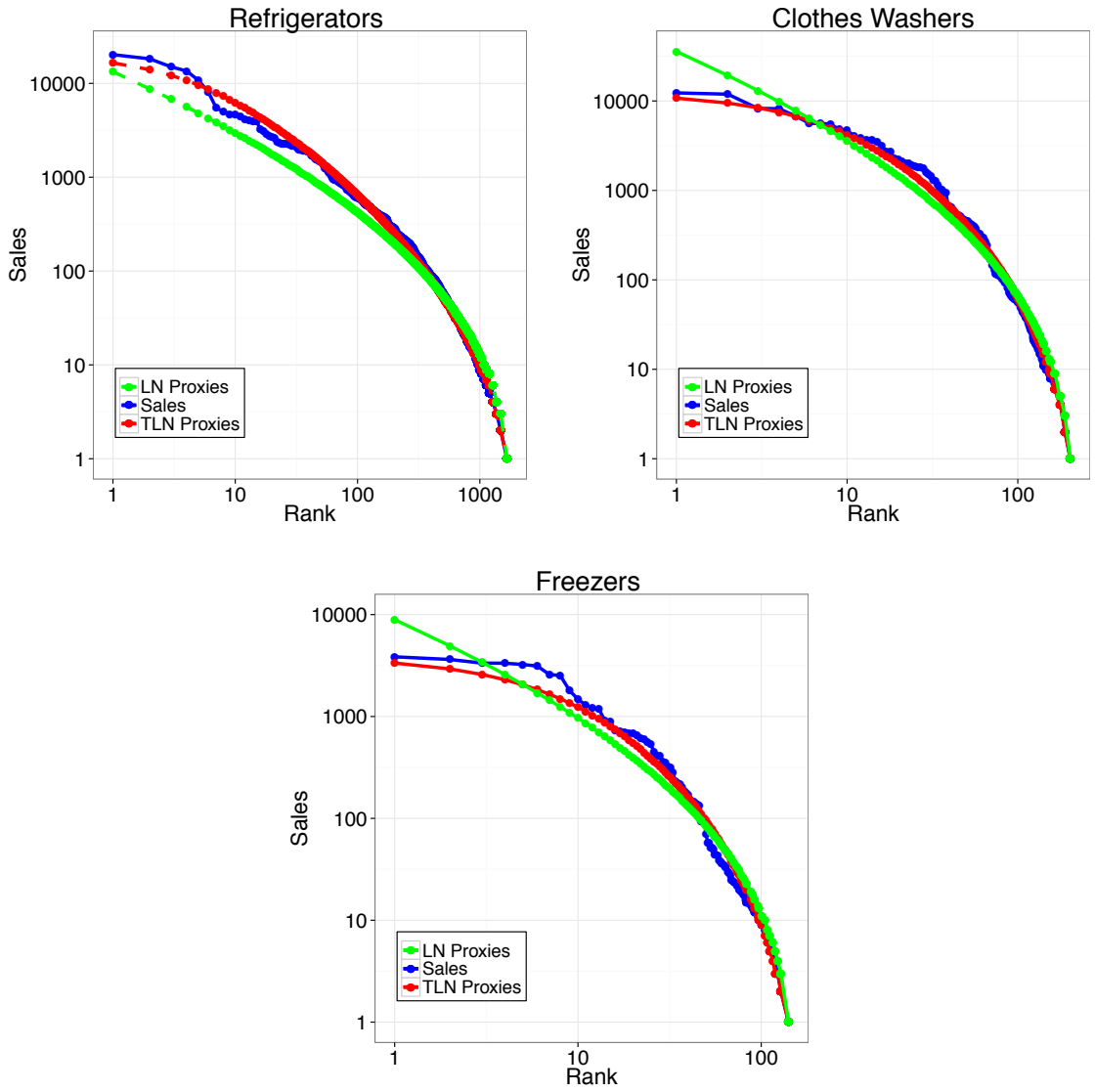


Figure 1: Zipf plot of sales rank versus sales for refrigerators, clothes washers, and freezers at the last time step.

Figure 1 displays the Zipf plots (Rank versus Sales) of the samples of the products along with the corresponding LN and TLN approximations. We observe that the patterns of the sales of the three products are represented by the LN and TLN to a reasonable degree of accuracy. Indeed, the R^2 values of the LN approximation and the TLN approximation are respectively: 0.981 and 0.998 for the refrigerators, 0.978 and 0.994 for the clothes washers, and 0.971 and 0.989 for the freezers. We also notice that the accuracy of the TLN approximation is superior to the LN, especially for the higher-ranked models of the freezers and clothes washers, where the LN proxies are overestimating the sales. Indeed, for clothes washers the LN sales estimate of the model function that corresponds to the first rank is equal to 35,823 units, which is an overestimation of 291% (the true actual value is 12,317), and the TLN approximation is 10,761, which is an underestimation of approximately 13%. For the freezer's first ranked model, the true value of sales is 3,852 and LN and TLN approximations of the sales are, respectively, 8,917 and 3,346, which represent respectively an overestimation of 313% and an underestimation of 17% respec.

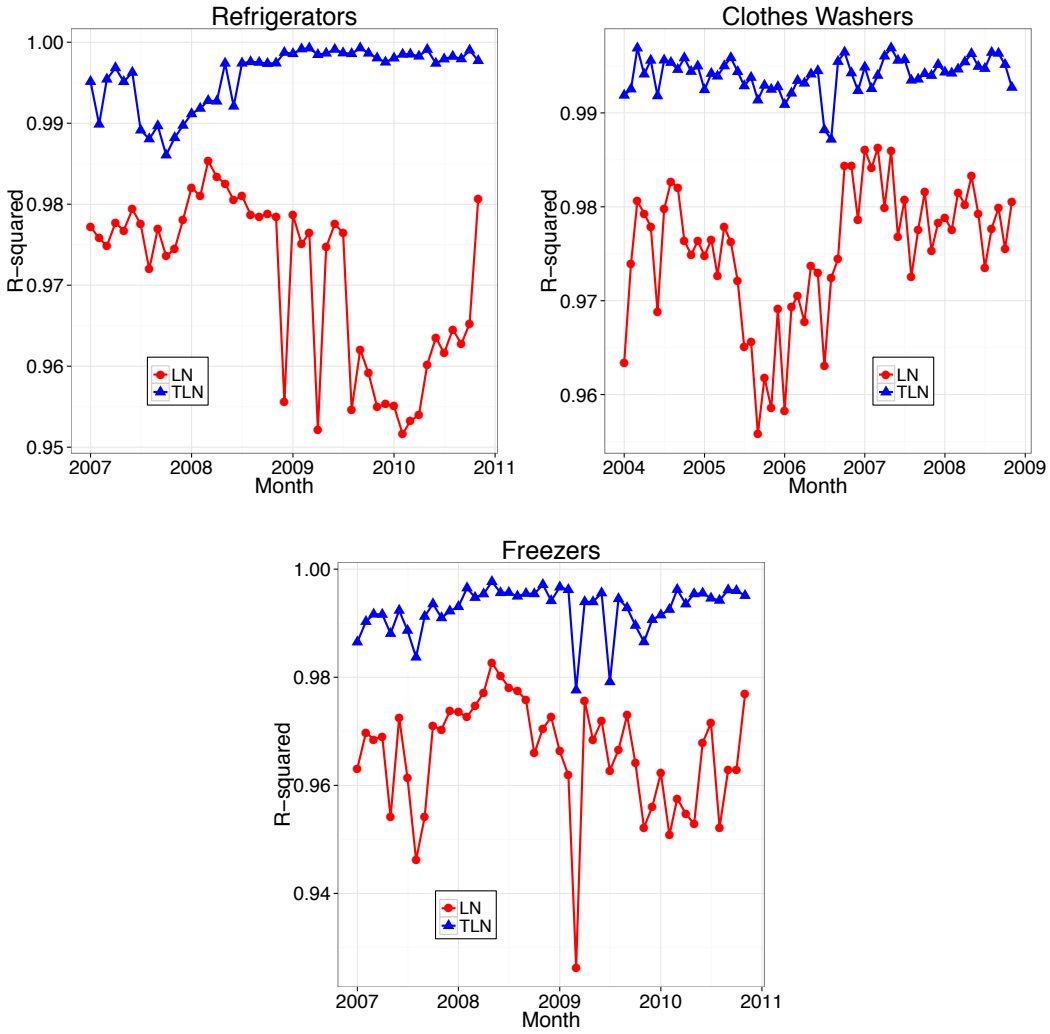


Figure 2: Values of R^2 for each time step of the learning period of the LN and TLN sales approximation of the refrigerators, freezers and clothes washers.

We then examined the approximation accuracy of the product sales during the learning period. Figure 2 displays the fitting quality index R^2 of the LN and TLN approximations for each time step covering the learning period. These results suggest that the LN and TLN approximations hold well, and it is clear that the TLN outperforms the LN for the three products. Table 1 summarizes the results of the calculation of the accuracy measurement. In addition to the R^2 values, we calculated the Mean Absolute Percentage Error (MAPE), which is an average of relative errors and it is defined as follow:

$$MAPE = \frac{100}{N}\sum_{i=1}^{N}\frac{|s_i - \hat{s}_i|}{s_i} \qquad (1.12)$$

where s_i and $\hat{s}_i$ are respectively the actual sales and model function sales estimate for rank r_i, and N is the number of models at the considered time step.

Thus, from Table 1 we note that TLN reduces the mean of the relative errors by 75% for refrigerators, by 58% for clothes washers, and by 65% for freezers.

	R^2		MAPE	
	LN	TLN	LN	TLN
Refrigerators	0.971 (0.010)	0.996 (0.004)	35.400 (9.591)	8.891 (5.732)
Clothes Washers	0.975 (0.007)	0.994 (0.002)	37.639 (7.940)	15.798 (3.322)
Freezers	0.966 (0.010)	0.993 (0.004)	46.557 (10.988)	16.070 (5.987)

Table 1: Mean values (and the standard deviation) of R^2 and MAPE for the pre-forecast period.

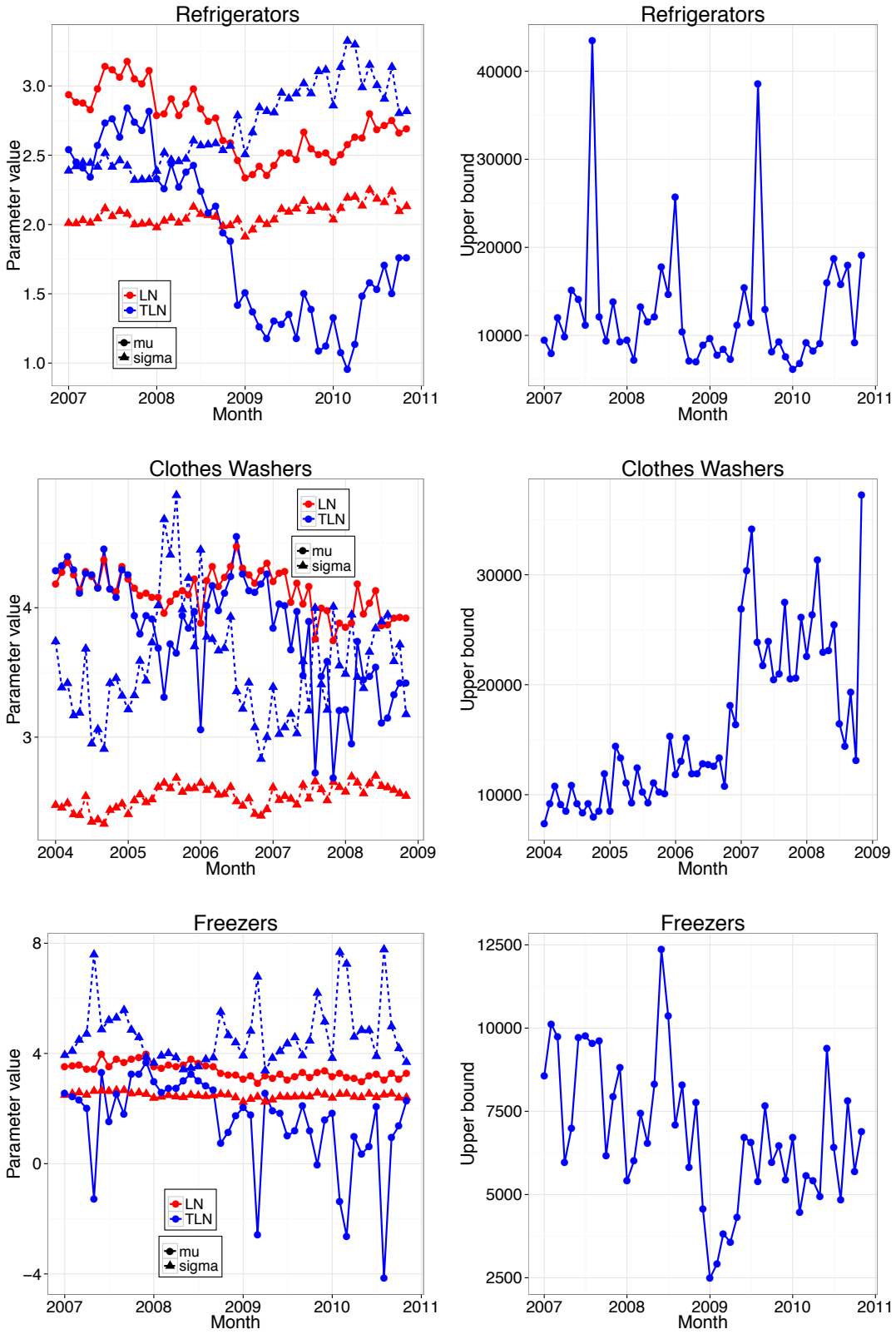


Figure 3: The estimated distribution parameters versus time.

The left column of the Figure 3 illustrates the evolution of the MLE estimates of the parameters $\bar{\mu}^{LN}$, $\bar{\sigma}^{LN}$, $\bar{\mu}^{TLN}$ and $\bar{\sigma}^{TLN}$ through the time steps. The dashed lines with triangles represent in red $\bar{\sigma}^{LN}$ and in blue $\bar{\sigma}^{TLN}$. The lines in red with circles represent $\bar{\mu}^{LN}$ and $\bar{\mu}^{TLN}$ in blue. We note that the TLN parameters have a high variability compared to those of LN. The right column of Figure 3 depicts the high variation of the upper bound ($\bar{b}$) through time. The mean values and the standard deviation of the parameter estimates ($\bar{\mu}^{LN}$, $\bar{\sigma}^{LN}$, $\bar{\mu}^{TLN}$ and $\bar{\sigma}^{TLN}$), as well as the upper bound ($\bar{b}$), are reported in Table 2.

	$\bar{\mu}^{LN}$	$\bar{\sigma}^{LN}$	$\bar{b}$	$\bar{\mu}^{TLN}$	$\bar{\sigma}^{TLN}$
Refrigerators	2.72 (0.04)	2.07 (0.03)	12690 (7273)	1.89 (0.11)	2.71 (0.08)
Freezers	3.38 (0.19)	2.48 (0.13)	6858 (2135)	1.57 (1.92)	4.69 (1.56)
Clothes Washers	4.15 (0.19)	2.52 (0.14)	13892 (6940)	3.86 (0.51)	3.69 (0.69)

Table 2: Mean values (and the standard deviation) of the distribution parameters.

5. Forecasting Sales

In the previous section we saw that the distribution parameter estimates vary significantly with time. However, we now empirically demonstrate that average estimates (as reported in Table 1) of the distribution function parameters ($\bar{\mu}^{LN}$, $\bar{\sigma}^{LN}$, $\bar{\mu}^{TLN}$ and $\bar{\sigma}^{TLN}$) can be used to forecast sales quantities using the sales ranks.

Given the estimates of the distribution parameters from pre-forecast data, we compute sales volumes using the sales rank for the forecasting period. Figure 4 shows that using these estimates can produce accurate results in terms of estimating future sales distributions. Table 3 summarizes the accuracy measurement for the sales distributions for the forecasting period. We note that the TLN outperforms the LN. Indeed, the mean relative errors are decreased by 70% for refrigerators, by 23% for clothes washers, and by 48% for freezers.

These results demonstrate that using the mean values of the distribution parameters can provide accurate approximations of sales using the sales ranking. It also shows that using the truncated version of the log-normal distribution improves the approximations.

	R^2		MAPE	
	LN	TLN	LN	TLN
Refrigerators	0.953 (0.007)	0.990 (0.011)	49.967 (6.144)	15.203 (11.010)
Clothes Washers	0.947 (0.013)	0.966 (0.011)	71.424 (11.762)	54.791 (11.565)
Freezers	0.953 (0.028)	0.979 (0.013)	58.286 (26.732)	30.161 (15.279)

Table 2: Mean values (and the standard deviation) of R^2 and MAPE for the forecast period.

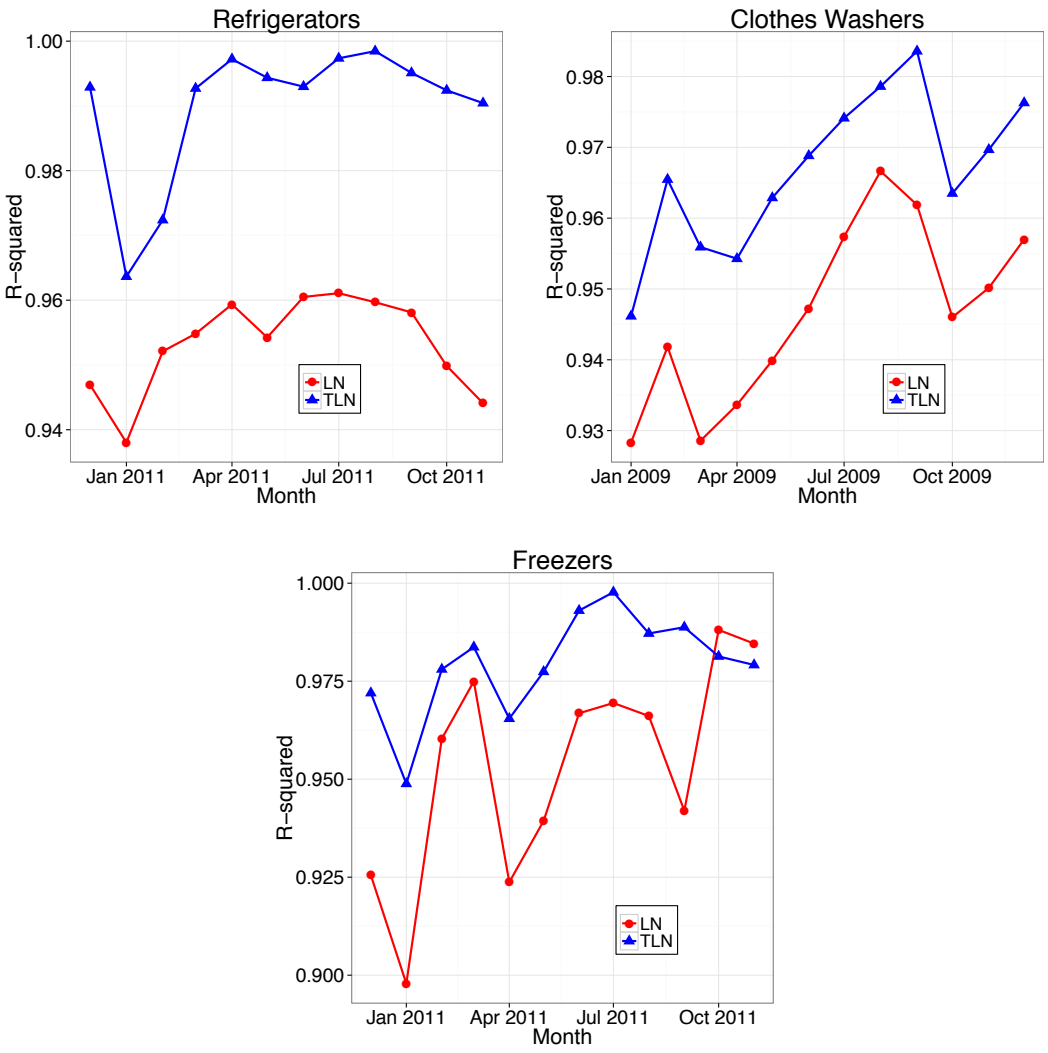


Figure 4: Values of R^2 for each time step of the forecasting period of the LN and TLN sales approximations for refrigerators, freezers, and clothes washers.

6. Application to Calculation of Market Average Quantities

In this section we focus our interest on computing two market average quantities: product price and product capacity. For all three products, each model has a volume capacity rating measured in cubic feet. For refrigerators and freezers, this attribute indicates how much food may be stored in the product, while, for clothes washers, this attribute is an indicator of what volume of clothes may be washed in the machine. These quantities are computed for each time step of the forecasting periods using four different data weighting methods.

The first method, which is the reference that represents the most accurate estimate, uses as a weight the true sales provided by the NPD Group's data. The second and the third methods use the sales proxies provided by the LN and TLN approximations. Finally, the fourth method weights the values for each product model equally. This last method may be used when an analyst has no estimates of sales or sales rank to provide weights for the model-specific data.

The left column of Figure 5 depicts the results of computing the product price average using the four weighting methods for the three considered products. We note that using the equally weighted price average for refrigerators and freezers produces a very high bias for the market average price estimate. We also note that, for the three products, the pattern of the evolution of the price averages through time using the equally weighted method misrepresents the reality, which can lead to a wrong interpretation of the results. For example, in October 2009 the equally weighted average price for clothes washers shows an increase, when the average price weighted by the true sales shows a decrease.

We now compare the MAPE measure to see how well the different methods of weighting approximate the price averages using the actual sales as a weight. The MAPE of the price averages for the methods using an equal weight, LN sales proxies weight, and TLN sales proxies weight are respectively 56.452, 1.512, and 1.665 for refrigerators; 11.103, 7.240, and 1.604 for clothes washers; and finally 116.077, 3.779, and 0.983 for freezers. The results for the equally weighted approximation confirm our previous observation, which is the presence of bias especially for refrigerators and freezers. For refrigerators, both sales proxies (LN and TLN) produce an accurate result, with a slightly higher accuracy for the LN sales proxies. For clothes washers and freezers, using the TLN sales proxies as a weight clearly outperforms using the LN sales proxies.

The right column of Figure 5 shows the results of computing the capacity averages using the four weighting methods through time. We can see that using the equally weighted approximation we lose all information about the variability of the capacity averages especially for refrigerators and freezers, and for clothes washers there is a high bias in the approximation. For freezers we note that the approximation using the LN sales proxies produces a high bias error as well. The MAPE of the capacity averages for the methods using an equal weight, LN sales proxies weight, and TLN sales proxies weight are respectively 2.230, 1.142, and 1.055 for refrigerators; 5.075, 1.689, and 0.214 for clothes washers; and finally 4.292, 4.658, and 0.672 for freezers.

These results shows that using the equally weighted market average quantities can produce misleading information about the products. It also shows that using the TLN distribution function to approximate the sales produces quite good results. However, the application of the fitted distribution function to estimating market average quantities requires the availability of sales rank data and historical data that enables the calculation of distribution function parameters.

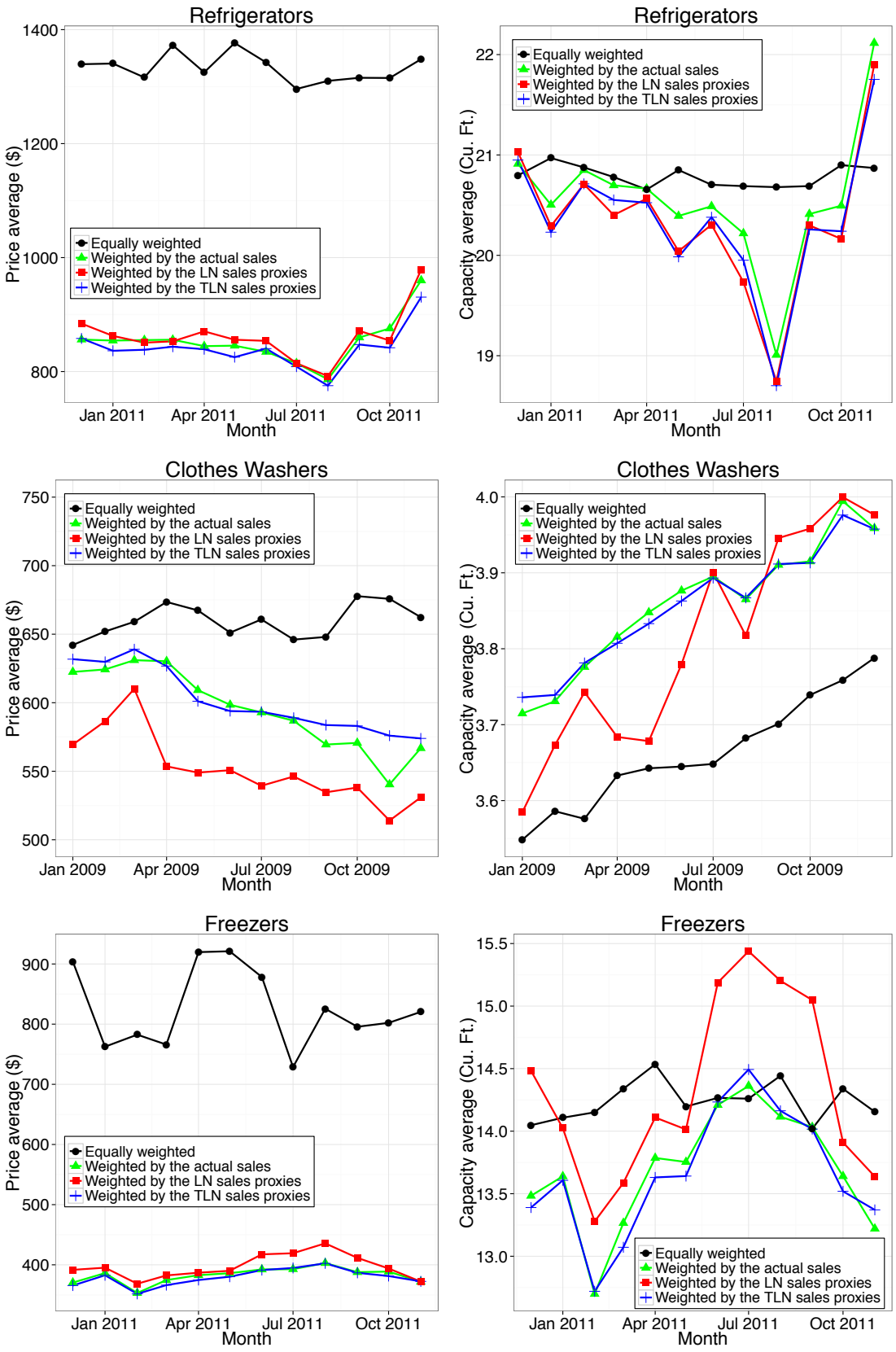


Figure 5: Price and capacity averages versus time using four different weighting methods.

7. Application to Calculation of Price Indices

As pointed out in Cavallo (2012), scraped data have the disadvantage that they do not provide information about sales quantities, which makes the calculation of the online weighted price indexes difficult. In this section we show that the availability of sales rank, and using the proposed method for estimating sales proxies, allows the calculation of a weighted version of the online price index. To compute the price indices of monthly price changes we use the same unweighted price index as in Cavallo (2013), which is a geometric average of price change, also known as the Jevons price index; we also use the Tornqvist price index, which is a weighted geometric average of price changes. The Jevons index is defined as follows:

$$P_{t,t-1}^{Jevons} = \left(\prod_{i=1}^{N_t} \frac{p_t^i}{p_{t-1}^i} \right)^{1/N_t}$$

where p^i_t is the price of the model i at time step t, and N_t is the number of models at time step t.

The Tornqvist index is defined as

$$P_{t,t-1}^{Tornqvist} = \left(\prod_{i=1}^{N_t} \frac{p_t^i}{p_{t-1}^i} \right)^{1/2 \left[\frac{p_i^t q_i^t}{\sum_{j=1}^{N_t} p_j^t q_j^t} + \frac{p_i^{t-1} q_i^{t-1}}{\sum_{j=1}^{N_{t-1}} p_j^{t-1} q_j^{t-1}} \right]}$$

where p^i_t is the price of the model i at time step t, q_i^t is the sales volume of model i at time step t, and N_t is the number of models at time step t.

To calculate these indices we used, at each time step, only the models for which price data can be observed in both time steps (t and t-1). Figure 6 shows the calculation results for the two price indices. The Tornqvist index is represented in three versions, which use in the weight calculation, respectively, the actual sales, the LN sales proxies, and the TLN sales proxies. We note that the Tornqvist indexes have much more variability than the Jevons, and they have different trends in some time steps. We also note that using the sales proxies provides a good estimation of the Tornqvist indexes. Indeed, if we consider the Tornqvist index with the actual sales as a reference, the MAPE measures for the Jevons indexes and the Tornqvist indexes using the LN proxies and the TLN proxies are: 1.700, 0.153, and 0.138 for refrigerators; 1.964, 0.657, and 0.333 for clothes washers; and finally 1.984, 0.546, and 0.129 for freezers.

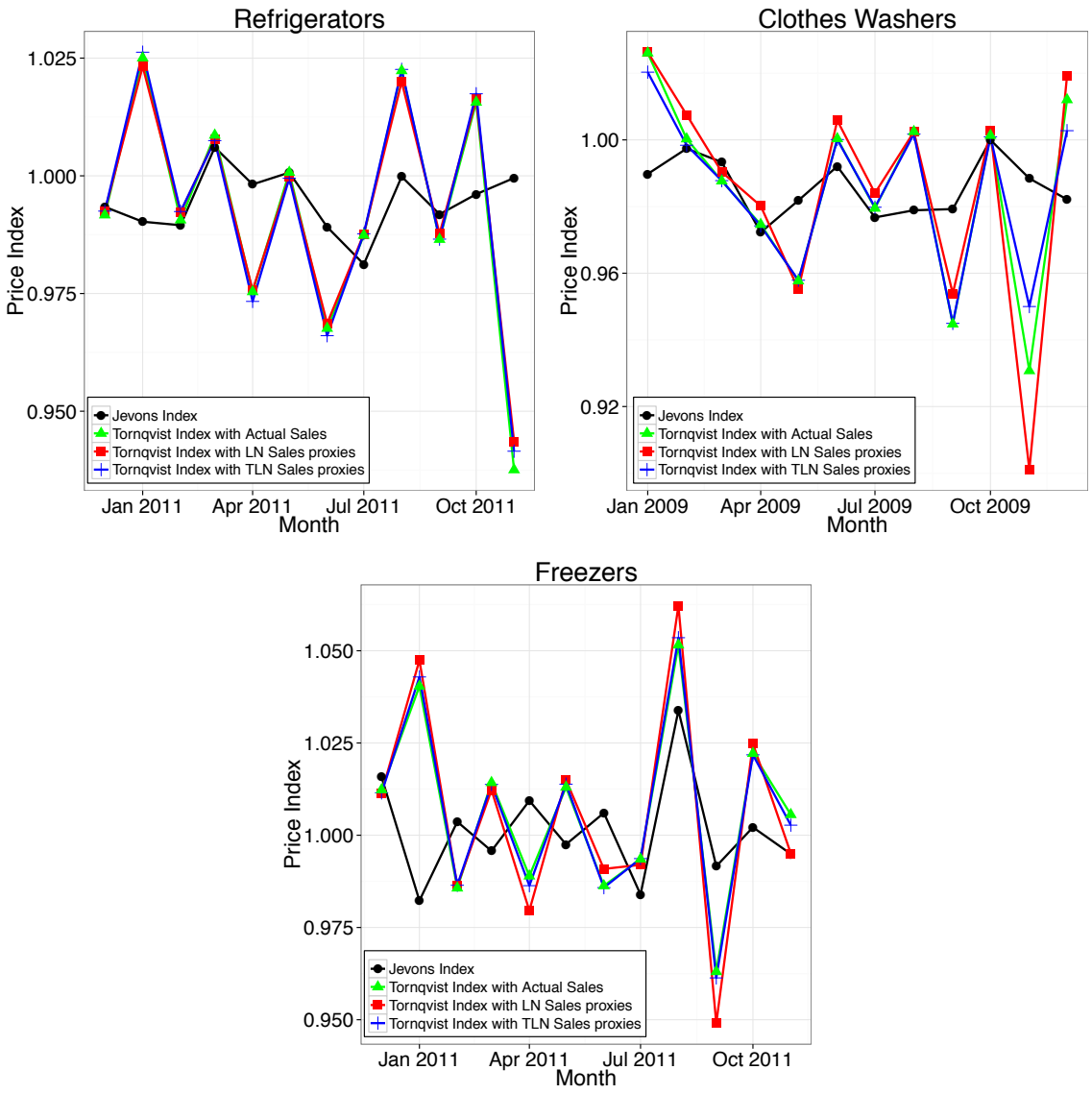


Figure 6: Price indices versus time using four different weighting methods.

8. Conclusion

In this work, we presented a straightforward method to produce an accurate approximation of sales volume using sales rank for refrigerators, freezers, and clothes washers. Our empirical results show that the log-normal distribution and specifically the truncated version are well suited to fit the sales distribution of the considered appliances. We also demonstrate the efficiency of this method, which results in a realistic estimation of product price averages and appliance capacity averages. We demonstrate that using sales proxies derived from a calibrated truncated log-normal distribution function produces realistic estimates of market average product prices, and product attributes during a forecast period where the distribution function parameters are assumed to not change over time. We show that the market averages calculated with the sales proxies derived from the calibrated, truncated log-normal distribution provide better market average estimates than sales proxies estimated with simpler distribution functions.

Acknowledgements

This work was supported by the U.S. Department of Energy under Lawrence Berkeley National Laboratory Contract No. DE-AC02-05CH11231. The authors are grateful to Dr. C. Anna Spurlock for her useful comments and constructive suggestions.

References

Cavallo, A. (2012). Scraped data and sticky prices.

Cavallo, A. (2013). Online and official price indexes: Measuring Argentina's inflation. *Journal of Monetary Economics 60* (2), 152–165.

Chang, C. H., Kayed, M., Girgis, M. R., Shaalan, K. F. (2006). A survey of web information extraction systems. *Knowledge and Data Engineering, IEEE Transactions on 18* (10), 1411–1428.

Chevalier, J., Goolsbee, A. (2003). Measuring prices and price competition online: Amazon.com and BarnesandNoble.com. *Quantitative Marketing and Economics 1* (2), 203–222.

Delignette-Muller, M. L., Pouillot, R., Denis, J.-B., Dutang, C. (2014). fitdistrplus: help to fit of a parametric distribution to non-censored or censored data. R package version 1.0-2.

Hisano, R., Mizuno, T. (2011). Sales distribution of consumer electronics. *Physica A: Statistical Mechanics and its Applications 390* (2), 309–318.

Johnson, N. L., Kotz, S., Balakrishnan, N. (1994). Continuous univariate distributions , vol. 1. John Wiley & Sons.

Mayer-Schönberger, V., Cukier, K. (2013). Big data: A revolution that will transform how we live, work, and think. *Houghton Mifflin Harcourt*.

Newman, M. E. (2005). Power laws, pareto distributions and zipf's law. *Contemporary Physics 46* (5), 323–351.

Pinto, C. M., Lopes, A. M., Machado, J. T. (2012). A review of power laws in real life phenomena. *Communications in Nonlinear Science and Numerical Simulation 17* (9), 3558 – 3578.

Spurlock, C. A. (2014). Appliance efficiency standards and price discrimination. *Lawrence Berkeley National Laboratory, Report Number* LBNL-6283E.

Stanley, M. H., Buldyrev, S. V., Havlin, S., Mantegna, R. N., Salinger, M. A., Eugene Stanley, H. (1995). Zipf plots and the size distribution of firms. *Economics letters 49* (4), 453–457.

Wasserman, L. (2004). All of statistics: a concise course in statistical inference. *Springer*.