Forecast UPC-Level FMCG Demand, Part I: Exploratory Analysis and Visualization

Dazhi Yang*, Gary S. W. Goh, Chi Xu and Allan N. Zhang
Singapore Institute of Manufacturing Technology (SIMTech)
*Agency for Science, Technology and Research (A*STAR)*
Singapore, Singapore
*Email: *yangdz@simtech.a-star.edu.sg*
yangdazhi.nus@gmail.com

Orkan Akcan
Antuit
Singapore, Singapore

***Abstract*—We are interested in forecasting a large collection of FMCG demand time series. As the demand of FMCG exists in a hierarchy (from manufacturers to distributors to retailers), the bottom level of the hierarchy may contain thousands or even millions of time series. Producing aggregate consistent forecasts while utilizing the unique features from each time series thus become a technical challenge. To achieve better forecasting results, exploratory analysis is often necessary to obtain insights on the underlying demand generating mechanism for each time series. Exploratory analysis aims at discovering those so-called "exogenous factors", such as price, demand of the complementary/substitutive goods and calendar events, which can help explain some of the demand fluctuation. During forecast accuracy evaluation, outlier detection is also important; a single anomalous time series can contribute much to the overall error. However, in a big data (such as retailing scanner data) enabled environment, exploratory analysis and visualization need much attention, because of the non-scalable nature of the existing methods. Scalability is essential for exogenous factor selection and outlier detection in big time series data. In Part I of this two-part paper, we introduce some exploratory analytics and visualization methods (from not scalable to very scalable) for big retailing time series. Forecasting of the hierarchical FMCG demand is addressed in Part II.**

***Keywords*-FMCG; forecasting; hierarchical reconciliation; visualization**

I. Introduction

Stock-outs have been an issue for fast moving consumer goods (FMCG) manufacturers, distributors and retailers for many decades [1]. Due to the non-durable nature of FMCG products, overstock situations are also not desired. Deteriorating products' prices are usually significantly reduced to stimulate demand, which directly translates to loss of potential profits [2]. In today's increasingly competitive retailing industry, enterprises seek to minimize, if not eliminate, the number of stock-outs and overstocks, as stock-outs directly affect consumer loyalty and overstocks translate to high inventory costs and wastage. Demand forecasting is thus essential for FMCG manufacturers, distributors and retailers to coordinate their efforts in the supply chain management processes to increase efficiency and improve consumer service levels [3].

Regardless of online or in-store retailing, the volume of demand-related data is enormous. As a result, retailing industry is one of the pioneers in using big data. The characteristics of the FMCG data align well with the HACE theorem[1] proposed in Ref. [4]. A typical retailer would collect scanner data over several dimensions including items, stores, markets, categories. These data give rise to the demand hierarchy.

There are many ways to segregate the FMCG demand. For example, a manufacturer produces various goods; these goods are shipped to various distributors and thus various retailers. In such context, the bottom level of the hierarchy contains the demand time series for each store and for each product. We aim to forecast such hierarchical FMCG demand in an aggregate consistent manner. In other words, the forecasts produced at lower levels should sum to the forecasts produced at higher levels. For example, at distributor level, the overall demand forecasts for a particular product are generated with a particular model. However, the demand forecasts for the same product but at store level are often generated with different models, as retailers are likely to have their own forecasting practice. Since the forecasts at retailer level are generated independently, they may not sum up to the distributor's forecasts. Therefore, the aggregate inconsistency in the forecasts becomes a technical challenge. We will address this issue in Part II of this two-part paper [5].

Given the size of a typical FMCG supply chain, the number of bottom-level series can easily reach an order of millions (e.g., 1000 retailers selling 1000 products). The big dimensionality of the data not only challenges

[1]Big Data starts with large-volume, heterogeneous, autonomous sources with distributed and decentralized control, and seeks to explore complex and evolving relationships among data [4].

forecasting, but also poses problems for exploratory analysis and visualization. For example, the complementary and substitutive effects in FMCG demand is well-known; by including these effects in forecasting models can improve forecast accuracy [6]. However, searching and identifying relevant exogenous predictors from millions of candidate predictors are not trivial.

Data inspection is a prerequisite to forecasting. Detecting anomalous time series in a big data context thus needs attention. In ordinary time series outlier detection, identifying anomalous data points from a univariate time series is the focus. However, when the data is a collection of time series, it is more amenable for forecasters to identify the unusual time series. Furthermore, when we consider data as hypercubes, i.e., each data point is associated with several tags (see Section II), missing data handling becomes crucial. List-wise deletion in this case may lead to unnecessary loss of data.

Based on the aforementioned challenges, three exploratory analytics/visualization methods are demonstrated. A tool for exploring/visualizing a single demand time series is provided in Section III-A. We found that by overlaying exogenous factors (such as price, promotion and calendar events) in a strategic way, the causal effects in FMCG demand can be studied. Section III-B introduces the so-called "kite plot", which is suitable for exploring/visualizing a moderately large number of time series. Kite plots are compact, allowing us to put many time series side by side, and thus visually identify similar time series. Furthermore, missing data can easily be represented within kite plot. Last but not the least, a scalable feature-based time series representation is shown in Section III-D. By summarizing the each time series into a set of predefined features and thus performing principal component decomposition, we can project big time series data onto a low-dimensional space. Anomalous time series, as well as the features contributing to the anomalies, can be identified through standard multi-dimensional outlier detection methods. Furthermore, when the time series is represented by features, we can assign appropriate forecasting models to a time series by examining the most influential features for that time series.

II. Data

We consider the Dominick's database[2] in this paper; the dataset is provided by the James M. Kilts Center, University of Chicago Booth School of Business with a collaborative effort by the Dominick's Finer Food (DFF). Although the dataset records weekly historical data from 1989 to 1994, owing to its informative nature, it is still frequently being used to conduct marketing research [7], [8]. It was previously found that sales in Dominick's database is large and frequent [9]; the dataset thus provides a suitable platform for our current investigation, i.e., forecasting under strong demand fluctuation.

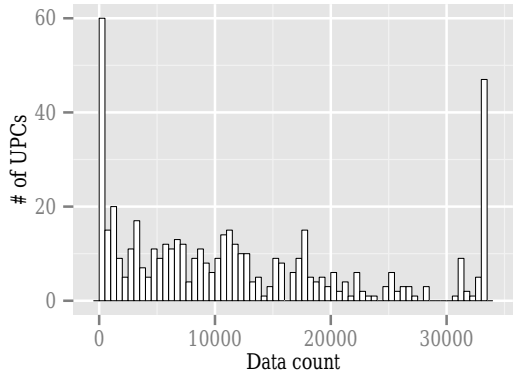

Figure 1. Histogram of UPC-level data availability for the bottle juice category.

The dataset contains four types of files, namely, the customer count file, store-level demographics file, universal product code (UPC) files and movement files. As detailed description of each file type can be found on the database website, we do not reiterate here. Instead we provide some insights on data preprocessing in the next section.

Store-level sales information for each UPC from 29 categories (such as beer, bottled juices, shampoos, etc) is recorded in the movement files. The data from each category can thus be viewed as data cubes (3-dimensional arrays) with dimensions: time ($\mathcal{T}$), store ($\mathcal{S}$) and UPC ($\mathcal{U}$). We consider three data cubes, namely, unit price, movement and promotion. As DFF will sometime bundle products (e.g., 3 cans of tomato soup for $2), the unit price of a product is obtained by dividing the raw price with quantity of the bundle. Movement reflects the number of product sold. Promotion indicates whether the product was sold on promotion in a particular week. There are three types of promotions, namely, bonus buy, coupon and simple price reduction. As most of the promotions are the bonus buy type, we do not distinguish types of promotion in this paper, as per Ref. [6].

During the data preprocessing, we found that data could be missing from any dimension, i.e., $\mathcal{T}$, $\mathcal{S}$ and/or $\mathcal{U}$. Furthermore, the missing data from each data cube may not locate at shared positions. Such nature of the data creates difficulties in comparing research results, as each individual researcher would apply different treatments for missing data. We therefore make our data processing code

[2]The dataset is freely available at http://research.chicagobooth.edu/kilts/marketing-databases/dominicks/.

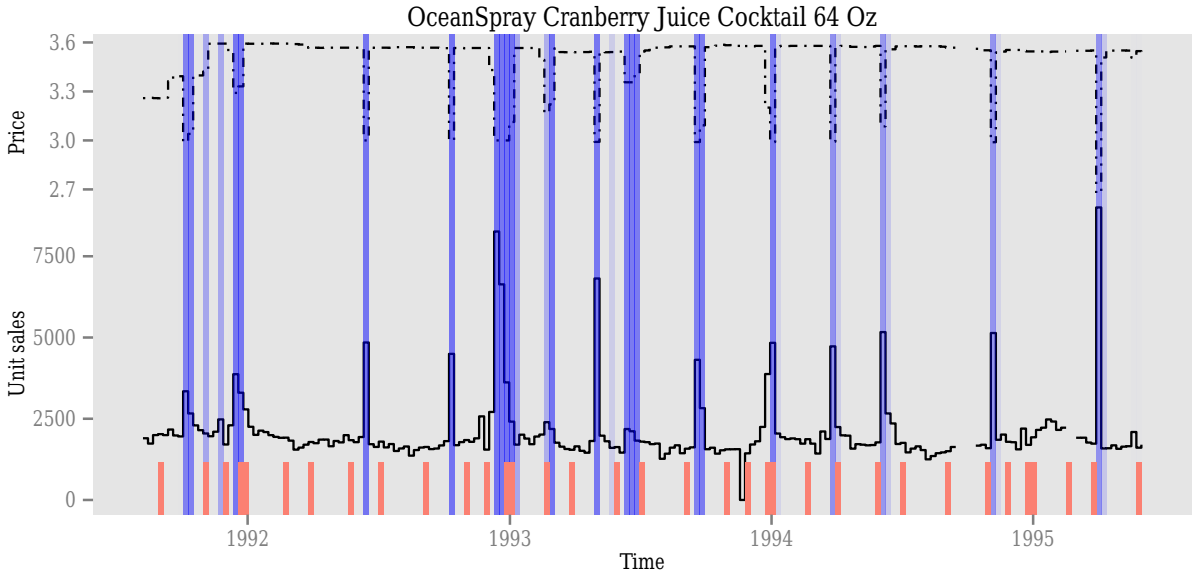

Figure 2. Unit sales and price transients of a particular UPC over a period from 1991 August to 1995 June. The promotional index is shown in blue (color opacity is proportional to promotional index value). Calendar events (9 public holidays in US) are shown in red.

available[3] and give a brief description on our particular preprocessing sequence.

Throughout this paper, only data from the bottled juice category (BJC) will be used for illustrative purposes. We acknowledge that forecast results from a single category do not represent those from the other categories. However, the focus here is to demonstrate the collaborative effect among distributors and retailers; the use of data from one category is considered sufficient. From the raw data files, the above-mentioned data cubes are constructed; the size of each data cube is $\mathcal{T} \times \mathcal{S} \times \mathcal{U}$, where $\mathcal{T} = 399$ (this is the case for all categories), $\mathcal{S} = 93$ and $\mathcal{U} = 511$. The total number of data points for each UPC is thus 37107 if no missing data is present. As shown in Fig. 1, out of the 511 UPCs in the BJC, only a handful of products have more than 30000 data points. We select the 40 UPCs with most data.

III. Exploratory Analysis and Visualization

The histogram filter used in Section II gives an overall idea about the data availability. However, to obtain a deeper understanding on the data and insights on the underlying demand generating mechanism, detailed visualization is required.

A. A Tool for Individual Time Series

As mentioned earlier, exogenous factors such as price and promotion information may help predict the demand. If the interest is to study a single time series, it is logical to overlay the exogenous factors together with the demand time series itself, so that any strong correlation can be detected visually.

In Fig. 2, the UPC-level sales of OceanSpray cranberry juice cocktail (1 of the 40 previously identified UPCs) over a period of ≈ 4 years is plotted using a solid black line. The store-wise averaged price of that UPC is shown by dotted line. The averaging weights are calculated based on all commodity volume (ACV) of each store[4]. All commodity volume is the total annual revenue of the store (available from the "PERregress" package in R [10]). The weighted average of the promotional index is calculated similarly[5]. The weekly promotional index is shown in Fig. 2 as the opacity of the blue stripes (less opaque indicates small index). Lastly, the calendar events are shown by the short orange bars at the bottom of the plot. It can be seen from Fig. 2 that a reduction in price (due to promotion) is an evident cause of the increase in sales. On the other hand, the seasonality (due to both yearly and calendar events cycles) in unit sales time series is weak. The effect of including

[3]Please contact the corresponding author for release of code.

[4]For example, suppose a product is being sold on 3.5 dollars with promotion in a store and the ACV of the store is 50 million dollars, and it is being sold on 2.5 dollars without promotion in another store with an ACV of 20 million dollars. The aggregated price would be $3.5\times(50/70) + 2.5\times(20/70) = 3.21$ dollar

[5]For the same ACV, the averaged promotional index is $1\times(50/70) + 0\times(20/70) = 0.71$, given $\mathcal{I} = 1$ if the item is on promotion in a store and 0 otherwise.

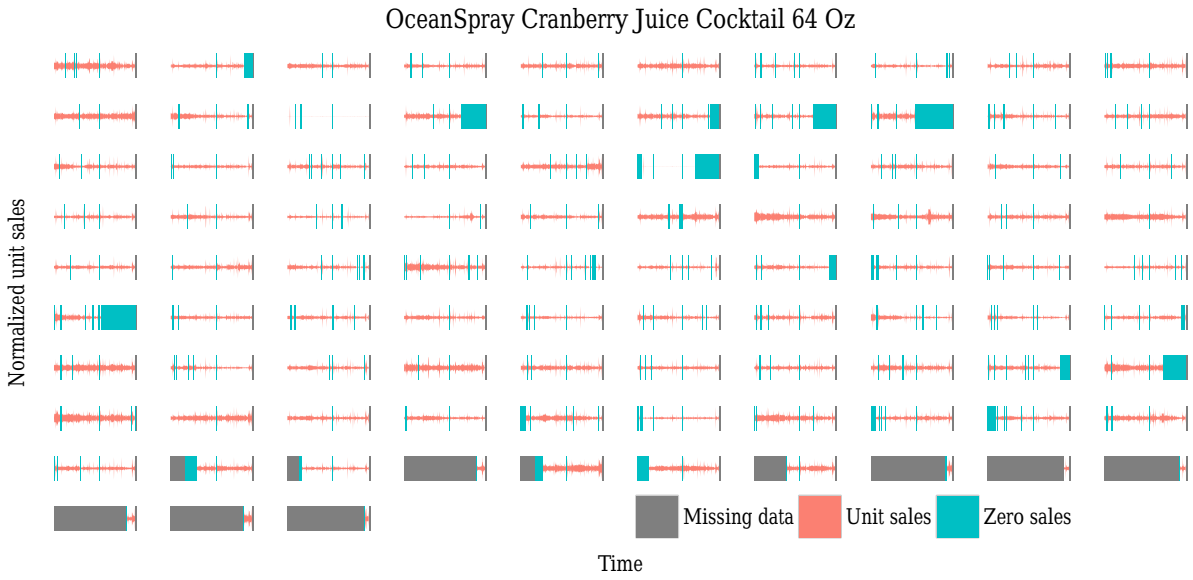

Figure 3. Kite diagrams for a particular UPC sold at 93 stores. The strictly non-negative store-level sales time series is flipped about x-axis to form closed glyphs. See text for detailed interpretation.

the calendar event and monthly dummy variables during forecasting is thus inferred to be insignificant.

We note that step functions are used in Fig. 2; such plotting design is essential for this visualization. As demand may not respond to the promotions and price changes immediately, plotting using step functions allows us to examine whether the promotions and price changes are aligned with demand. Given a long enough time series, without such plotting design, it is impossible to identify any misalignment.

B. A Tool for Many Time Series

Plots like Fig. 2 consider a single or a few time series at a time. If the number of time series gets large, the resultant "spaghetti" is not informative. Therefore, we can employ a so-called "kite diagram" [11] to visualize the time series. A kite diagram uses closed, symmetric glyphs to represent data. In this case, the glyphs are formed by flipping the strictly non-negative sales series about the x-axis. The space-filling representation supports time series transient visualization for a moderately large number of time series. In additional, markers can be used to highlight unusual data values such as missing data and zero sales. Fig. 3 shows an example kite diagram for store-level demand of OceanSpray cranberry juice cocktail. The plot contains rich information about each time series; it also helps identify time series with similar transient, which is very likely to be useful during forecasting. Some other observations that can be made from Fig. 3 include store commodity values (reflected by the width of the glyphs), promotional effect (relative magnitude of the sales spikes) and whether the product is being sold at a particular store (for long consecutive zero sales, the product is most likely not being sold anymore, or if the store is not yet open for business).

Although the kite plot is more compact than a simple time series plot, it is still not very scalable. Arguably we can insert more time series into the plot, however, as the number of time series gets over a few hundred, visual identification of similar time series becomes difficult.

C. Data Preprocessing Sequence

Before we introduce the scalable visualization tool for big time series, based on previous visualizations, data preprocessing sequence is first described. The processed data will be subsequently used in Part II of this two-part paper for forecasting.

After the histogram filter used in Section II, the total number of store-level time series is 3720 (93×40). We thus have three matrices of size 3720×399 for unit sales, price and promotion. Based on the observations, a preprocessing sequence is designed for the DFF data:

1) Fig. 3 shows that the data from last few weeks are constantly missing across stores. A list-wise deletion (for all three matrices) is performed if there are more than 2000 missing values in any column (weeks) of any matrix. The size of the matrices is 3720×387 after this step.

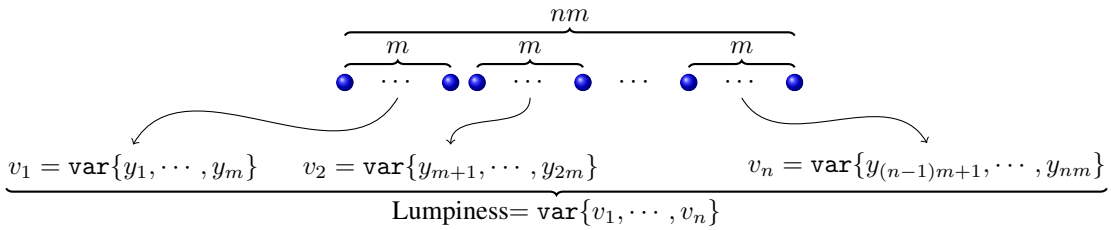

Figure 4. Calculation of lumpiness for a time series of length nm. Symbol `var` denotes variance.

2) After step 1, it is observed that most time series do not contain any missing value. We therefore remove the rows (stores) with any missing value based on the unit sales matrix. The size of the matrices is 3036×387 after this step.
3) Fig. 3 indicates (by consecutive zero sales) that a UPC may no longer be sold at a particular store. Time series with more than 10 consecutive zero-sales weeks are therefore filtered. The size of the matrices is 2409×387 after this step.
4) It is observed that zero sales may occur across the stores for a particular UPC for a particular week. As our aim is to consider hierarchical nature of the sales, we remove columns (weeks) if zero sales are spanning any UPC. The size of the matrices is 2409×377 after this step.
5) We repeat step 2 with respect to the price matrix, i.e., delete the time series with any missing price. The size of the matrices is 1991×377 after this step.
6) Spurious unit sales spikes are also observed in the data. We consider a data point as a spike if it is 300 times more than the mean of the time series; those time series containing any spike are removed. The final dimensions of the matrices are 1971×377.

As our application is forecasting, our preprocessing is designed to preserve the temporal structure as much as possible (by sacrificing the number of bottom-level time series). We note that the above preprocessing is ad hoc but thorough. It also illustrates certain preprocessing difficulties for a typical dataset.

D. A Tool for Big Time Series

In this section, we demonstrate a scalable time series visualization technique using those 1971 time series obtained after the preprocessing. Instead of considering the time series space, we consider a feature space. A set of d predefined features, see Table I for example, is extracted from each time series. In other words, each time series becomes a unit of observation in the feature space and is described by the d-dimensional coordinates. Principal component analysis is then applied to the features. Finally, the exploratory visualization can be carried out through the classic biplot [12].

Table I
VARIOUS TIME SERIES FEATURES USED IN THIS PAPER. WE NOTE THAT THESE FEATURES ARE ADOPTED FROM REF. [13].

Feature	Description
ACF1	First order of autocorrelation.
Trend	Strength of trend.
Linearity	Strength of linearity.
Curvature	Strength of curvature
Entropy	Spectral entropy.
Lumpiness	Changing variance.
Spikiness	Strength of spikiness.
Lshift	Level shift using rolling window.
Vchange	Variance change.
Fspots	Flat spots using discretization.
Cpoints	The number of crossing points.
KLscore	Kullback-Leibler score.
Change.idx	Index of the maximum KL score.

A total of 13 features ($d = 13$) are defined in Table I. Some of these features are intuitive while the others may not be straightforward to understand. For example, lumpiness is defined as the variance of the variances of data segments from fix-sized time windows. To calculate lumpiness, the raw time series is first divided into n length-m data segments, see Fig. 4. The variance of each data segment is calculated, e.g., $v_1 = \texttt{var}\{y_1, \cdots, y_m\}$; and lumpiness is given by $\texttt{var}\{v_1, \cdots, v_n\}$. For a more detailed description on various features shown in Table I, we refer the readers to Ref. [13] and the references therein. Nevertheless, the set of features is non-exhaustive, other features could be defined based on applications.

After the features are calculated, PCA is performed. Fig. 5 shows the biplot of the 1971 time series. Each number in the plot represents one time series; a point along a particular vector is best represented by that vector. For examples, series-1787 at the top right corner is found to be spiky and lumpy, where as series-1219 on the left has strong first order autocorrelation.

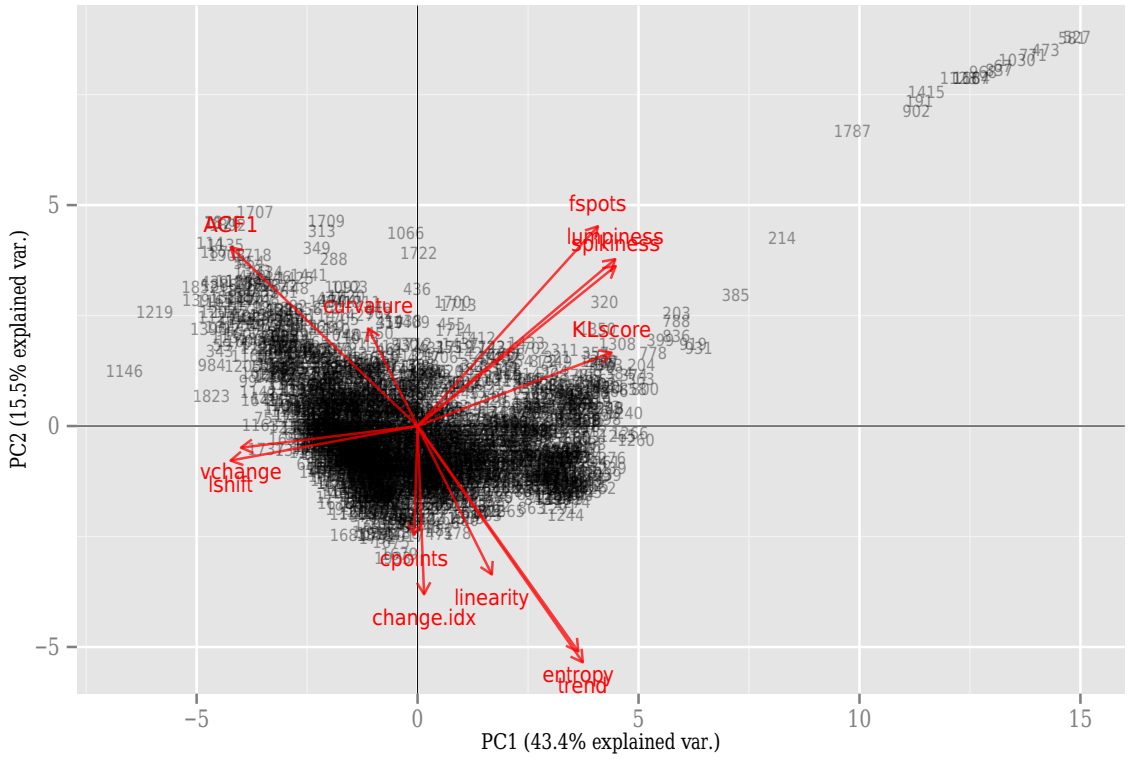

Figure 5. Biplot of the 1971 time series (represented by the numbers). 13 features are extracted from each time series and used in the principal component analysis.

IV. Discussion on the Principal Componential Based Visualization

As the first few principal components are usually sufficient to represent most variability in the feature space (as indicated in Fig. 5, the first two principle components explains about 60% of the variability of data in the feature space), this algorithm can efficiently identify the anomalous time series with respect to all other time series in the collection. Once the bipolar is produced, standard multi-dimensional outlier detection algorithms such as the highest density region method [14] and α-hull method [15] can directly be applied.

Another potential application of the principal component based visualization is to help identify suitable forecasting models. Using the earlier example, series-1787 is spiky and lumpy, a variance stabilizing step, such as the square root transform, prior to forecasting may be useful. Series-1219 has strong autocorrelation, time series models such as the autoregressive model may be appropriate. We will consider the forecasting model tailoring in a future paper.

Beside the wide applicability, another distinct advantage of the method is its scalability. Although only 1971 time series are used in the present study, the visualization is capable to handle much bigger dataset (Fig. 5 will contain more points, that's all). However, the feature computation step may be time consuming. The total run time for calculating the features for all 1971 series is 62 s on a late 2013 MacBook Pro computer. If we calculate features for a million time series, approximately 8 h is needed. Nevertheless, the forecast horizon for FMCG is usually week-ahead, the computation time is thus not an issue in operational forecasts.

Before we end this section, we would like to note that the PCA-based method can directly operate on the raw time series, i.e., each data point in the time series becomes a feature. However, it is not recommended due to the following two reasons: (1) the results are difficult to interpret. Unlike Fig. 5 where the properties of each time series can be easily associated with the feature names, if raw time series is used, this would not be possible. (2) the lengths of the time series need to be identical, which is almost impossible to obtain in reality. On the other hand, if features are considered, the length of the time series becomes less important. We only need to ensure a same

number of features are extracted from each time series.

V. Conclusion

Due to the hierarchical nature of the demand (from manufacturers to distributors to retailers), big time series data often exist in manufacturing supply chains. Several exploratory analytics and visualization tools are proposed in this paper to help discover useful knowledge and information embedded in the data.

When the interest is in a single time series, we recommend to overlay exogenous factors (such as price and promotion information) together with the demand time series, so that the correlation among the factors and demand can be visually inspected. When dealing with a moderately large number of time series, we recommend to use the kite plot. Similarities among various time series and the data availability can be examined. Furthermore, some time series characteristics such as the variability can also be derived from the kite plot. Finally, for big time series data, the principal component representation is suitable. Each time series is reduced to a set of predefined features. The properties of each time series, as well as the anomalies can be summarized and derived from the standard biplot.

In Part II of this paper, some of the knowledge discovered through the exploratory analysis herein shown is utilized in forecasting. More specifically, we are interested in producing forecasts for hierarchical FMCG demand in an aggregate consistent manner.

Acknowledgment

This work is partially supported under the A*STAR TSRP fund 1424200021 and Antuit–SIMTech Supply Chain Analytics Lab.

References

[1] W. E. Wecker, "Predicting demand from sales data in the presence of stockouts," *Management Science*, vol. 24, no. 10, pp. 1043–1054, 1978.

[2] A. Chande, S. Dhekane, N. Hemachandra, and N. Rangaraj, "Perishable inventory management and dynamic pricing using RFID technology," *Sadhana*, vol. 30, no. 2-3, pp. 445–462, 2005.

[3] F. R. Jacobs and R. B. Chase, *Operations and Supply Chain Management*. Mcgraw Hill, 2010.

[4] X. Wu, X. Zhu, G.-Q. Wu, and W. Ding, "Data mining with big data," *Knowledge and Data Engineering, IEEE Transactions on*, vol. 26, no. 1, pp. 97–107, Jan. 2014.

[5] D. Yang, G. S. W. Goh, S. Jiang, and A. N. S. Zhang, "Forecast UPC-level FMCG demand, Part II: Hierarchical reconciliation," in *Big Data (Big Data), 2015 IEEE International Conference on*, Oct 2015.

[6] T. Huang, R. Fildes, and D. Soopramanien, "The value of competitive information in forecasting FMCG retail product sales and the variable selection problem," *European Journal of Operational Research*, vol. 237, no. 2, pp. 738–748, 2014.

[7] R. C. M. Daniel Toro-González, Jill J. McCluskey, "Beer snobs do exist: Estimation of beer demand by type," *Journal of Agricultural and Resource Economics*, vol. 39, no. 2, pp. 174–187, 2014.

[8] A. Jami and H. Mishra, "Downsizing and supersizing: How changes in product attributes influence consumer preferences," *Journal of Behavioral Decision Making*, vol. 27, no. 4, pp. 301–315, 2014.

[9] R. A. Chahrour, "Sales and price spikes in retail scanner data," *Economics Letters*, vol. 110, no. 2, pp. 143–146, Feb. 2011.

[10] P. Rossi., *PERregress: Regression Functions and Datasets*, 2013, r package version 1.0-8. [Online]. Available: http://CRAN.R-project.org/package=PERregress

[11] S. Thakur and T.-M. Rhyne, "Data Vases: 2d and 3d Plots for Visualizing Multiple Time Series," in *Advances in Visual Computing*, ser. Lecture Notes in Computer Science, G. Bebis, R. Boyle, B. Parvin, D. Koracin, Y. Kuno, J. Wang, R. Pajarola, P. Lindstrom, A. Hinkenjann, M. L. Encarnacao, C. T. Silva, and D. Coming, Eds. Springer Berlin Heidelberg, 2009, no. 5876, pp. 929–938.

[12] K. R. Gabriel, "The biplot graphic display of matrices with application to principal component analysis," *Biometrika*, vol. 58, no. 3, pp. 453–467, 1971.

[13] R. J. Hyndman, E. Wang, and N. Laptev, "Large-scale unusual time series detection," 2015, working paper.

[14] R. J. Hyndman, "Computing and graphing highest density regions," *The American Statistician*, vol. 50, no. 2, pp. pp. 120–126, 1996.

[15] B. Pateiro-Lopez and A. Rodriguez-Casal., *alphahull: Generalization of the Convex Hull of a Sample of Points in the Plane*, 2015, r package version 2.0. [Online]. Available: http://CRAN.R-project.org/package=alphahull