Coping with demand volatility in retail pharmacies with the aid of big data exploration

Papanagnou, C and Matthews-Amune, O

http://dx.doi.org/10.1016/j.cor.2017.08.009

Coping with Demand Volatility in Retail Pharmacies with the aid of Big Data Exploration

Christos I. Papanagnou

Salford Business School
University of Salford, Manchester, M5 4WT, UK

Omeiza Matthews-Amune

Salford Business School
University of Salford, Manchester, M5 4WT, UK

Abstract

Data management tools and analytics have provided managers with the opportunity to contemplate inventory performance as an ongoing activity by no longer examining only data agglomerated from ERP systems, but also, considering internet information derived from customers' online buying behaviour. The realisation of this complex relationship has increased interest in business intelligence through data and text mining of structured, semi-structured and unstructured data, commonly referred to as "big data" to uncover underlying patterns which might explain customer behaviour and improve the response to demand volatility. This paper explores how sales structured data can be used in conjunction with non-structured customer data to improve inventory management either in terms of forecasting or treating some inventory as "top-selling" based on specific customer tendency to acquire more information through the internet. A medical condition is considered - namely pain - by examining 129 weeks of sales data regarding analgesics and information seeking data by customers through Google, online newspapers and YouTube. In order to facilitate our study we consider a VARX model with non-structured data as exogenous to obtain the best estimation and we perform tests against several univariate models in terms of best fit performance and forecasting.

Keywords: Retail Pharmacy, Data Mining, Time Series, Forecasting, Big Data, Demand Uncertainty

1. Introduction

Due to the increasing prominence of concomitant complex factors regarding consumer choices, retail industry has been challenging with demand volatility and uncertainty, which affects demand planning and inventory holding (Azadeh et al., 2015). The detrimental consequences of this uncertainty in demand are often amplified from downstream echelons (retail stores) to upstream echelons (manufacturing) in a phenomenon referred to as bullwhip effect (Lee et al., 1997; Papanagnou and Halikias, 2008). Another major consequence of demand volatility is the increasing inaccuracy

of forecasts which have resulted in excessive stocking leading to expiries and losses especially when considering products with a predetermined shelf life (Betts, 2014). Literature suggests a considerable amount of various univariate time series models and common estimation techniques in drug retail to address demand volatility. However, these cannot tackle limitations related to their applicability on historical data from short past windows, which are often characterised by linearity (see Buzia et al., 2016; Anusha et al., 2014). Demand planning and inventory control in the retail industry have long been adversely affected by the combination of a complex set of factors determining consumer choices (Chao, 2015). Consumer behaviour related factors ranging from social, cultural and economic to psychological and personal have been implicated by researchers as complicit in promoting demand volatility, which can bring negative consequences on inventory control, consumer confidence and shareholder profits in the long run (Ferreira et al., 2015). Considering the high stakes involved with retail demand prediction, it is a vital problem for every retail company to address fluctuations on medicine inventories. The pharmaceutical retail business presents a unique challenge as profits are increasingly threatened by short expiry dates, increasing government regulation in the sale of medicines and fierce (sometimes monopolistic) competition by rivals (Gibson, 2012).

Cadeaux and Dubelaar (2012) attributed the challenge of demand volatility faced by retail pharmacies to the lack of a reliable inventory management system which should provide useful forecasting information on relationships between sales volatility and stock at the level of the specific product item. This has resulted in overstocking in some cases leading to excessive inventories, while understocking on the other hand resulting in poor customer satisfaction levels (Mahar et al., 2012). Yadav (2015) argue that retail pharmacies are a popular choice in low-income countries for individuals seeking healthcare for minor ailments as a result of the ease of access as compared to the bureaucratic processes, cost and time involved in hospital visitations. Also, in many smaller towns where hospitals are unavailable or reside in bigger cities, retail pharmacies are the first point of call for treatment and advice. Anusha et al. (2014) identify stock outs due to poor demand planning whereas the high cost of medicines arising from increased length of the supply chain. Considering the importance of retail pharmacies in the health care supply chain, accurate demand planning is critical to balancing demand and supply, which ensures product availability, waste reduction, customer satisfaction, improved inventory management, minimisation of over and under stocking and increased profits.

Linoff and Berry (2011) elucidate that accurate demand estimation is critical to minimising over and under stocking thereby minimising losses and most importantly maximises sales and customer satisfaction. Liu et al. (2013) attempt to address this challenge by both proposing a simulation optimisation system for pharmacy inventory management using empirical distributions to model demand. In a simulation-based approach, Kroger Company - which operates 1,950 in-store pharmacies around the United States - managed to reduce out-of-stocks but no consideration was taken in terms of demand seasonality, trends and "disturbances" in demand pattern (Zhang et al., 2014).

Watson et al. (2014) compared the efficacy of various six sigma analytical models in improving the inventory replenishment policies and minimising out of stock rates in a large volume pharmacy setting. While out of stock rates were reduced, the research was based on process improvement and undesired "shortage" events with a lack of consideration of historical demand data. Ribeiro et al. (2016) adopted Pegels method a multiplicative trend exponential smoothing based technique - in order to guarantee sufficient stock levels and perform better short-term forecasting horizons for specific drug types in a pharmaceutical distribution company by investing solely on demand patterns. However, the presence of demand fluctuations constitutes a constraint for this approach as the proposed model can become very sensitive to volatility. In addition, the authors argue that the validity of the results is subject to additional data such as data types considered in the present paper.

Other studies suggest that the pharmaceutical retail demand may be challenged from a different angle by considering temporal and economical features that are associated with consumers and may affect demand in an extreme gradient boosting model (Sawon et al., 2016). However, the main limitation of this approach revolves around the lack of use of customer related data in favour to store related data. One solution for this could be given by merging customer related data in order to minimise drug inventory levels in a large medical distribution organisation. Neural networks computing systems were used to merge very basic customer related information (e.g., gender, unique customer number), which was obtained from already existing transactional (but limited) data in the store (Bansal et al., 1998). In another study, a combined ARIMA methodology with artificial neural networks (ANNs) was advanced in order to capture with accuracy both linear and nonlinear patterns of sales in a drug retail store (Khalil Zadeh et al., 2014). In particular, an explorative network based analysis was conducted on medicine sales by analysing linear covariations due to the lack of enough past sales records for each drug and customer related data.

The importance of considering customer related data by focusing on the impact of new channels of information sourced from the internet on consumer purchasing habits has been explicated in pharmaceutical industry. With the aid of internet-based platforms - like social media applications which encourage consumer interactions - the internet has become a primary health-related information source allowing data propagation through user-generated content and sharing capabilities (Shankar and Li, 2014; Tyrawski and DeAndrea, 2015; Greene and Kesselheim, 2010). Wosinska (2002) also documents the rising influence of internet-based direct-to-consumer advertising on supporting patients choices for several medicines and treatment. This has resulted in an increase in the total market demand for a therapeutic group as well as increased purchase frequency by means of greater therapy compliance.

Considering the rising importance and the vital role that consumer generated data from the internet plays on drug purchasing preferences, Kim et al. (2015) underscore the shortage of relevant existing demand planning research as the majority of available studies focus on simple data analysis. For the impact of such complex internet data to be evaluated, Thomassey and Fiordaliso (2006) suggest that data mining and machine learning have been shown to provide better results than statistical models in nonlinear data structures or when complex relationships exist. Hamuro et al. (1998) also support

the notion that harnessing consumer information from data mining provides retail pharmacies with a competitive advantage in the demand estimation process.

To clearly assess the interdependencies of complex relationships obtained by data mining of the internet consumer-generated data, several VAR models have been suggested (Khalil Zadeh et al., 2014). The advantages and applicability of VAR models rest on the fact that they allow all key variables in the model to be considered symmetrically by identifying an equation for each variable, which contains lags and delays of all other variables in the model. Despite their numerous benefits, VAR models have been applied only scantily for demand estimation in pharmaceutical industry. Watson et al. (2014) favour the application of advanced VAR models in minimising out of stock rates in pharmacies by emphasising that the assumptions of conventional inventory theories (e.g., EOQ policies) were insufficient and unrealistic. Kim et al. (2015) attempted to evaluate the impact of user generated data from social media blogs by employing the VARX model in considering the impact of exogenous variables on upstream demand forecasting on pharmaceutical supply chains. They showed that the strengths of the VARX model reinforced by its ability to simultaneously analyse the impact of all the variables in the system on each other - making it highly adaptable to structural changes.

This paper proposes the use of the VAR in the analysis of the impact of exogenous variables obtained from a variety of internet based sources of endogenous historic sales data. According to Kenny (2011), exogenous variables are determined by external factors outside the model, which have an impact on the endogenous variables. This study suggests that Google search intensity, online newspaper keywords and YouTube video duration, which are treated as external variables, can form the exogenous variables. These variables can be explanatory as information seeking from patients may influence demand for medicines (Suziedelyte, 2012). The inclusion of external variables in estimation and forecasting models has increased in recent times as a result of the realisation that external explanatory factors - which have an impact on other variables existing within the model - may affect accuracy. Employing this model is expected to address the deficiencies of other "orthodox" models discussed by evaluating the complex nature of relationships and interdependencies in systems were situations like nonlinearity and seasonality coexist with a critical emphasis on the impact of consumer generated data on real business estimating situations.

1.1. Big data analytics and demand estimation techniques

The increase in quantity and speed of universal data generation from various sources over the last few decades has resulted in the emergence of big data term. Sources of such data have been identified as digital videos, texts and pictures, social media sites, online business transactions, climate sensors etc. However, questions have arisen on how non-structured data related to specific products can be used to improve inventory management - including departmental merchandising decisions - and how managers can evaluate what would happen under uncommon circumstances derived by generalised and multiple-sourced customer-based estimators of seasonality. In supply chain context, companies may benefit from big data analytics by exploring how sales data can be used with detailed customer sentiment data to improve inventory management

either in terms of forecasting or treating some inventory as committed based on specific shoppers requirements. Laney (2001) initially identified three characteristics of big data as Volume, Variety and Velocity. McNulty (2014) suggested an addition of Veracity and Value to the other three characteristics bringing it to a total of "5Vs".

The application of advanced analytics to the enhancement of visibility across the entire supply chain is cogitated by a considerable amount of organisations (Hazen et al., 2014; Waller and Fawcett, 2013; Lee et al., 2013). Applying supply chain analytics to the analysis of complex data sets provides managers with the ability to respond to relevant problems in a timely manner by perceiving accurate business insights. Product range in retail industry varies from a wide range of general consumer goods to electronics to medicines. With items like medicines - especially nonprescription drugs which are subject to demand fluctuation as a result of internal factors (e.g., promotions, sales events, pricing) and external factors (e.g., seasonality, duration of ailment, competition and weather) - predicting the demand accurately can be very challenging (Jain et al., 2014).

Demand estimation modelling takes into account external and internal factors to retail pharmacies as this would help in streamlining and optimisation of business decisions. Depending on if these external and internal factors meet the "5Vs" criteria, they could be classified under the broad term big data. Comparing the 2010 allergy season to the same period in 2015, Chao (2015) identifies the use of big data in advanced supply chain analytics as responsible for the improvement in demand forecasting leading to drastically reduced stock outs of allergy-relief products in northeastern U.S. retail pharmacies. Khalil Zadeh et al. (2014) identified forecasting weaknesses as the major cause of excessive inventories and perverse drug shortages. Last, there are also numerous reports in retail pharmacy chain regarding improvements in operational planning efficiency and significant savings with the aid of advanced big data analytics (see Wang et al., 2016; Jones and Gupta, 2015).

Utilising data analytics not only strengthens supply chains but also increases the chances of timely customer order fulfilment subsequently ensuring brand loyalty and higher profit margins as this research envisages (Fahimnia et al., 2015). As the potential for big data analytics to provide companies with competitive advantage increases, it is critical for managers in a retail pharmacy to understand information flow and how to generate business intelligence out of data. Information gathering from diverse sources can provide such business intelligence systems with the ability to support operational planning through a wide range of functions. To support this quest for business intelligence knowledge, this research aims to study the impact of big data through text mining on forecasting accuracy for pharmaceuticals using Nigeria based HMX pharmacy stores as a case study. HMX pharmacy stores located in Abuja, and it is classified as an SME. HMX operates as a pharmaceutical retail outlet providing a wide range of OTC medicament and prescription only medications (POM) as well as other household consumables. Other services provided include doctor prescription fill-ups and pharmaceutical care.

We trust that big data is a relatively new term and it is often associated with the rapid proliferation of data generated by humans, machines and nature in the form of digital processes. In some occasions, it is believed that big data relates mainly to

extremely large data sets or large pools of data, and as a result, the emphasis is given to the Volume (Fiore et al., 2013; Ohlhorst, 2012; Chen et al., 2012). In this research - we follow the fashion of many practitioners and academics - and we espouse the idea of big data exploration across all "5Vs". In particular, Variety is represented by the diverse data (structured, semi-structured and unstructured) sources, the amount of data (129 weeks) - which could be scaled up - signifies the Volume of data collected, while Velocity is characterised by the speed of obtaining this data at retailer's end. Value refers to the importance of these data in determining the impact of exogenous variables on forecasting accuracy while Veracity signifies the reliability of the data obtained from trusted secondary sources.

2. Types of data

Considering the challenges posed by the increasing availability of new information sources characterised by velocity, volume, variety which may influence consumer demand resulting in demand volatility, it is clear that the conventional forecasting techniques discussed in currently available literature may be inadequate to account for the influence of these exogenous variables. Hence, a need exists for research which considers the impact of these exogenous variables on forecasting accuracy. This paucity of relevant research validates further the importance of this study. In this research we consider data from nonprescription medicines, which can be given over the counter (OTC). Canadian consumer health products (2012) defines nonprescription medicines as drugs "used for the prevention, treatment, symptomatic relief, cure or risk reduction of diseases, injuries or chronic conditions which are sold off shelves and do not require a doctors prescriptions" Historically, estimation and forecasting techniques have been applied extensively to predict the demand for OTC medicines with the aid of linear and Poisson regression models, several univariate time-series and hybrid neural networks (Ribeiro et al., 2016; Kirian and Weintraub, 2010; Socan et al., 2012).

2.1. Structured data

The structured data for this study is generated by demand figures, which in this study is represented by the quantity of weekly sales of OTC medicine (PM) from January 2014 to June 2016 (129 weeks). PM is an analgesic available for a wide range of indications and more details are given in section 3. Note, that the use of medicine sales data has been broadly applied in the past mainly to predict demand patterns in retail pharmaceutical industry (see for example, Anusha et al., 2014; Riahi et al., 2013).

2.2. Semi-structured data

This is computed by taking the average of the Google index and Newspaper Keyword Index (NKI). Such combination of news and web search index data has been applied by Sun et al. (2015) who predicted the fluctuation of real estate market by combining online news articles data and web search data and Nardo et al. (2016) who multiplied Google trend and Yahoo finance news data to obtain a unified index to forecast weekly stock price changes. The keywords in this study are obtained from

Google *AdWords* website, which provides the volume of searched keywords relating to the indication of the medicine in the particular region. Based on search volumes, Table 1 shows the top ten searched keywords related to keyword "pain" for PM drug used in this study.

Table 1: Top ten searched keywords for PM drug

Drug type	**Keywords**
Analgesic	Pains, Fever, Aches, Malaria, Pain Relief,
	Migraine, Neck Pain, Tooth Aches, Dysmenorrhea, Arthritis

Apart from longitudinal demand data, this paper opts for logical and systematic collection of semi-structured and unstructured information from specific preexisting sources (Online Newspapers, Google Trend website, YouTube and sales data). In most cases, these types of (secondary) data would not require reexamination as they are characterised by a certain degree of reliability and validity based on their sources. Furthermore, since the data window analysis is specific and can be adjusted accordingly there is a low level of inaccuracy or the danger to analyse outdated data. However, the format of semi-structured and unstructured is often a stumbling block for direct analysis of data, which is a very common problem for big data analytics (Power, 2014). Thanks to transmutation methods non-structured data can be available in formats which may be easier and less expensive for the researcher to assess. As researchers have little control over data quality, aggregations and definitions may be unsuitable. Secondary data can also be described as raw data if already collected and summarised but is yet to be processed, analysed or compiled. Note that our approach may include higher volumes of data (by simply encountering data from longer time windows), which can be processed by MapReduce and Hadoop Distributed File System (HDFS), both inextricably intertwined with Hadoop (Dean and Ghemawat, 2008).

2.2.1. Weekly Google index

Many researchers have attested to the predictive power of Google searches in different sectors in the past using Google index (Askitas and Zimmermann, 2009; Li et al., 2015; Tuhkuri et al., 2016). The weekly Google index *GI* - representing Google searches for keywords related to the word pain - is obtained from the Google trends website, which stores weekly search data since 2004. Data sets from Google trends have been applied as exogenous variables for demand forecasting (see Choi and Varian, 2012; Pan et al., 2012; Bratina and Faganel, 2008). In principle, data from Google trends represents volume ratios of keywords searched in specific geographical areas related to the word pain instead of absolute volumes. More specifically, Google trend provides the number of times a keyword had been searched as a fraction of the total number of Google search queries at the same period. The weekly *GI* is then computed by first aggregating the weekly Google trend values of all keywords had been searched for a particular (predefined) area and then by dividing each aggregated weekly figure by the sum of all weekly figures for the period of study. The figure obtained can be

then multiplied by 100 for better interpretation reasons. The formula for obtaining the Google index $GI_{t,i}$ at week t in a geographical area i, is shown in (1).

$$GI_{t,i} = \left[\frac{\frac{S_{t,i}}{R_{t,i}}}{\max_t(\frac{S_{t,i}}{R_{t,i}})} \right] \times 100, \quad t \in [1, N_w] \tag{1}$$

where $S_{t,i}$ represents the number of searches with keywords $k_{t,i}$ and $R_{t,i}$ denotes the total number of search queries in geographical area i at each week t in total N_w weeks. It should be noted that the value of $\frac{k_{t,i}}{GI_{t,i}}$ is the keyword search ratio (also referred to as search volume index), which provides rising searches related to a query and can be also obtained directly from Google trends. In case where Google searches occur in multiple geographical regions the weekly Google index $GI_{t,i}$ can be simply aggregated as $GI_{t,a} = \sum_{i=1}^{a} GI_{t,i}$, where a is the total number of different geographical areas.

2.2.2. Weekly Newspaper Keyword Index

The utilisation and analysis of news content data in prediction has been applied by researchers in the past. Kholodilin et al. (2014) applied a quantitative content analysis technique based on word count, wherein the impact of keywords is determined quantitatively by computing a keyword index. Word count involves counting the frequency of keywords with the assumption that those words appearing most often indicate important concerns. However, the downside of the word count is the use of synonyms, which may underestimate the importance of concepts and multiple meanings of words and, consequently, may mislead the researchers. For a single keyword, Kholodilin et al. (2014) obtained the monthly scaled word by dividing the number of monthly occurrences of the keyword by a proxy for the overall monthly text volume. In case more than one key topics were involved, Kim et al. (2015) performed a topic trend analysis by averaging disease-related topic weights occurring in a social media blog over the same month to obtain a common monthly weighted average.

As multiple keywords exist in this study, similar approach is applied in this paper to obtain first the weight of each keyword obtained from Google *AdWords* used earlier, and to calculate then the unified weekly keyword index by averaging all keyword weekly weights. To obtain the individual keyword weight, all news articles from the selected newspapers containing the keywords in a single week are extracted. Afterwards, the frequency of each keyword, for each newspaper and for each day is divided by the total word count of the weekly aggregated articles, where keywords were found, to give a weight of the particular keywords during that week. The weekly keyword index (WKI_t) at week t is obtained by (2) where f_i represents the number of occurrences (frequency) for each keyword at day i. $k_{j,n}$ represents each keyword j at newspaper n and $TW_{i,n}$ denotes the aggregated total word count of articles at day i in each newspaper n, where each of the keyword j was found. Last, N_D, N_K and N_N represent the number of days, keywords and newspaper, respectively.

$$WKI_t = \frac{\sum_{i=1}^{N_D}\sum_{j=1}^{N_K}\sum_{n=1}^{N_N} f_i k_{j,n}}{\sum_{i=1}^{N_D}\sum_{n=1}^{N_N} TW_{i,n}}, \quad t \in [1, N_w] \tag{2}$$

2.3. Unstructured data

A very common unstructured data types are the digital video formats. Gandomi and Haider (2015) argue that although video analytics is still in its infancy as compared to other big data sources, it could be applied in collecting key demographic information from customers. Analytics can also detect the unique and repeated visits as well as the amount of time spent in particular store areas, which can help in predicting buying behavior and assist in demand forecasting for different demographic groups shopping within the retail store. In another study, Verbeke and Ward (2001) assessed the impact of mad cow disease by investigating Belgian demand for fresh meat from 1995-1998 using media index of TV coverage and advertising expenditures as explanatory variables. Kim et al. (2015) used social media data from a health blogging website as the exogenous variable in a VARX model to forecast demand for medicines.

In this study, we consider unstructured data obtained from YouTube videos. YouTube is the third most visited website after Facebook and Google (Ericsson consumer lab, 2015). The utilisation of YouTube video data in determining the most popular and viral videos has been investigated by researchers (Figueiredo, 2013; Borghol et al., 2011). Barfar and Padmanabhan (2015) employed conventional TV programme viewership data in predicting the outcomes of the 2012 US elections. In particular, they computed the minutes per voter watching the TV programmes in a specific location during the run up. This approach, can be also used to obtain the number of views per minute of viewers watching specific YouTube video related to the pain keyword for a pre-specified period T. To obtain the number of views per minute, the weekly videos - relating to relevant keywords (the same approach followed previously for semi-structured data) - are consolidated, and the weekly video duration (minutes) is divided by the aggregated number of weekly YouTube views. The minutes per viewer MPV_t at week t is obtained by equation (3) where VD_t is the total video duration at week t and TV_t is the total number of views at the same weekly period. This leads to the following formula:

$$MPV_t = \frac{\sum_j VD_t}{\sum_j TV_t}, \quad t \in [1, N_w] \tag{3}$$

3. Model specification

Literature suggests that univariate methods might be not adequate to cope with demand uncertainty (Bratina and Faganel, 2008; Jones et al., 2009). The advent of internet and pervasive technologies allow customers nowadays to access sources that are appropriate with their decision-making on when and what products to buy. Thus, companies are facing a new challenge, which comes up with the analysis of additional

data in the same breath with other external factors (e.g., data from social media or advertising campaigns). In many cases, however, a feedback between endogenous and exogenous variables does exist, such that both variables may affect each other. In this study, a feedback exists between both variables as information seeking from Google, newspapers and YouTube, which collectively form the exogenous variables, can affect the demand for medicines (endogenous variable), while the ownership of medicines can also lead to information seeking. The existence of the relationship between information seeking by consumers for health services and its impact on demand behaviour was explored by (Suziedelyte, 2012). Multivariate autoregressive models can offer to a great extend the analysis of the level of relationship between these variables and can capture linear interdependencies among multiple time series models (Kim et al., 2015; Carrizosa et al., 2016).

In our study, we assume the following VARX*(p,s)* model:

$$Y_t = \alpha + \sum_{j=1}^{p} \Phi_j Y_{t-j} + \sum_{j=0}^{s} \Theta_j^* X_{t-j} + \varepsilon_t \tag{4}$$

where $X_t = (X_1t, X_2t, ..., X_rt)'$ is an r-dimensional time series vector representing exogenous input variables to the predictor variable Y_t, Θ_j^* is a $k \times r$ matrix and ε_t is a vector white noise process with $\varepsilon_t = (\varepsilon_1t, \varepsilon_2t, ..., \varepsilon_rt)'$ with mean vector $E(\varepsilon_t) = 0$ and covariance matrix $\Sigma = E(\varepsilon_t\varepsilon_t')$. For a more detailed description of the vector autoregressive process with exogenous variables and interpretation of the coefficients in the model (see for example, Box et al., 2015).

3.1. Data Collection

To carry out this study, weekly historical sales (demand) data from January 2014 to May 2016 for a specific drug (class) type was collected. The drug class is paracetamol (acetaminophen), abbreviated to PM, which is classified as a wide-range analgesic and because it is the most widely available and used painkiller worldwide due to an absence of significant contraindications (Obu et al., 2012). Other analgesics such as NSAIDS (ibuprofen, diclofenac) and those containing codeine were considered but not selected due to their contraindications in ulcer and hypertension, respectively.

Now suppose that predictor variable Y_t in (4) is partitioned into three groups of subcomponents with dimensions k_1, k_2 and k_3, respectively such that $k = k_1 + k_2 + k_3$ and $Y_t = (Y_1t, Y_2t, Y_3t)'$. In a similar way, the vector noise process can be partitioned as $\varepsilon_t = (\varepsilon_1t, \varepsilon_2t, \varepsilon_3t)'$. Thus, initially we will assume that in terms of model specification, Y_t is a 3-dimensional vector time series for which a VAR model needs to be estimated. Then, we assume that semi-structured and unstructured data can represent inputs to the system, while demand variables can denote the output. Consequently, we let $X_t = (Y_1t, Y_2t)'$ a vector with exogenous variables to the system that influence the output - represented by structured data - Y_3t. Finally, we define as $Y_1t = GI_{t,i} + WKI_t$ the aggregated semi-structured data and with $Y_2t = MPV_t$ the unstructured data as described in section (2).

4. Example

In this section, we show the model building procedures for OTC drug type. The time series considered consists of T=129 observations, on the three variables Y_1t, Y_2t, and Y_3t. Data for demand time series Y_3t was multiplied by *0.1* to ease the analysis. Figure 1 shows clearly a trend pattern as demand increases from the beginning of the year and peaks towards April, then falls and increases again in June, falls around November and then picks up again in December. The first high demand period coincides with the annual harmattan season in Nigeria beginning from December till around March. Air pollution during harmattan season is caused by the poor quality of air, which is regularly characterised by hot and dry wind laden with dust from the Sahara Desert. This causes often respiratory diseases resulting in flu-like symptoms like headaches, which can be remedied by OTC drug intake. Sales again pick up around June and begins to fall between September and October which coincides with the annual raining season in Nigeria which runs from May/June to September/October. This rainy season promotes rapid mosquito population development and parasite maturation leading to malaria in adults and children. The annual world malaria day comes up on 25th of April, at the beginning of the rainy season, and it is marked with free malaria screening activities and loads of mass media adverts to increase awareness and this may also contribute to the increase in sales occurring in subsequent months.

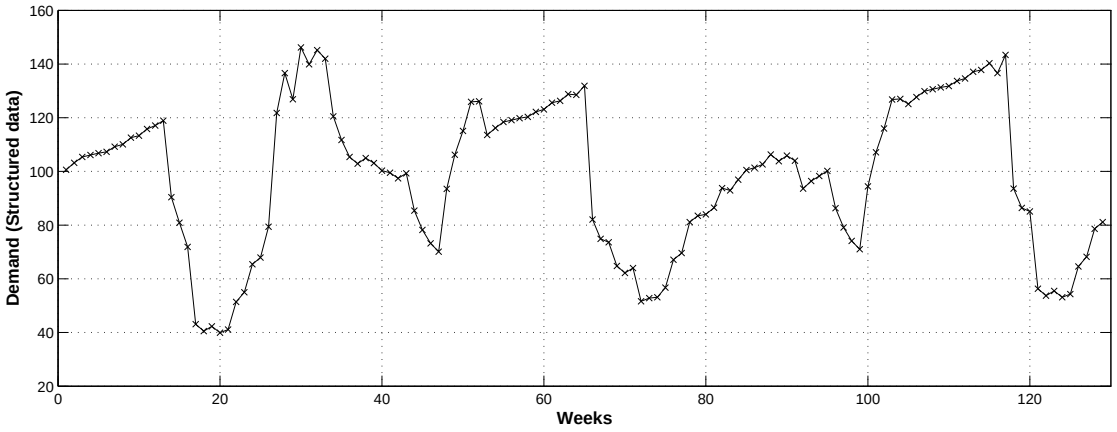

Figure 1: Y_3t, Weekly Demand Data Series for OTC drug (Analgesic)

It is also obvious from the plot that the highest demand for the OTC drug observed between June and August 2014, which coincided with the Ebola disease outbreak of 2014 (July to October) in Nigeria. Fever, headaches and muscle aches are symptoms of Ebola fever and these symptoms can be treated with analgesics. Our newspaper content analysis also showed that the word Ebola was mentioned 56 times in online newspapers between July to October 2014 compared to 13 times in the same period in 2015. This increased media coverage may have led an increased number of people to have medical tests when they exhibited symptoms like headaches and fever and this may have translated to a higher demand for acetaminophen. The highest demand for information related to pain and headaches is also shown in the spikes between July and

October 2014 of unified Google and newspaper keyword index shown in Figure 2 and YouTube minutes per view plot illustrated in Figure 3. This period also coincides with the Ebola outbreak in Nigeria. The increases in information seeking from mass media, Google and YouTube is also fairly consistent with the seasonal pattern of spikes also observed during the rainy season (May/June to September/October) and harmattan season (December to March).

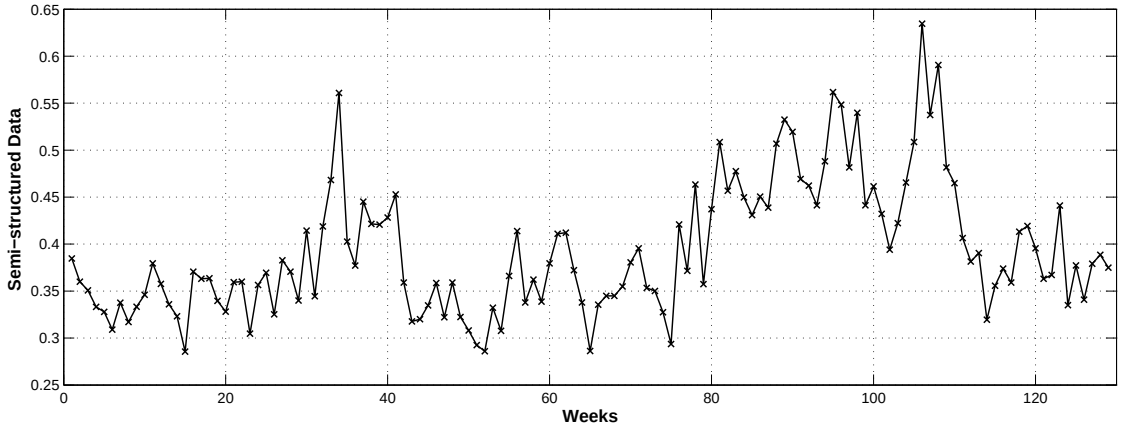

Figure 2: Y_1t, Weekly Data Series for Google Index GI and Newspaper Index, WKI

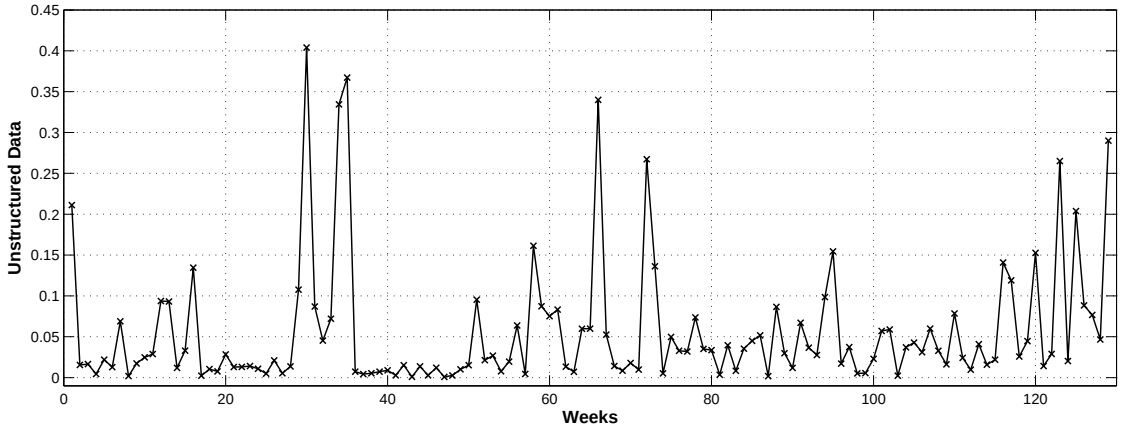

Figure 3: Y_2t, Weekly Data Series for YouTube minutes per view, MPV

5. Results and Discussion

First, we attempted to estimate different VAR models of orders $m = 1, ..., 7$ to the 3-variable series Y_t by least-squares. The significance of each m-th order VAR matrix was tested by $H_0 : \Phi_m = 0$ against $\Phi_m \neq 0$, in case when a VAR(m) model has been fitted to the series. The tests were performed with the aid of likelihood-ratio (LR) testing principle by using the following statistics (see Reinsel, 2003, p.

95) $\lambda_m = (mk + 3/2 - T + m)\log(U_m)$, where $U_m = \frac{|S_m|}{|S_{m-1}|}$, $S_m = \tilde{\Sigma}_m \times T$ is the Maximum Likelihood (ML) estimator of an order m model and S_{m-1} is the sum of squared residuals for $\Phi_m = 0$. In order to identify the best order for the VAR model, we opt for three selection criteria: Akaike Information Criterion (AIC), Hannan and Quinn Information Criterion (HQIC) and Schwarz's Bayesian Information Criterion (SBIC). Table 2 summarises the results from fitting the VAR model.

Table 2: Statistical results from fitting AR models to demand data for OTC drug

m (VAR Order)	1	2	3	4	5	6	7
$\lvert\tilde{\Sigma}_m\rvert(\times 10^3)$	1.1492	1.0027	0.9434	0.7025	0.6612	0.6240	0.5458
λ_m	360.7086	16.3018	7.0302	32.8833	6.5072	5.9859	13.3272
AIC_m	-6.5868	-6.6216	-6.5374	-6.6848	-6.5956	-6.5012	-6.4804
$SBIC_m$	-6.6294	-6.6243	-6.5415	-6.6903	-6.6023	-6.5092	-6.4897
$HQIC_m$	-6.6946	-6.7560	-6.7408	-6.9585	-6.9409	-6.9194	-6.9728

Results in Table 2 show the an AR model of order four can offer the best fitting among the other AR models. The LS estimates from the AR(4) model are:

$$\hat{\Phi}_1 = \begin{bmatrix} 0.5764 & -0.1201 & -0.0002 \\ -0.0835 & 0.2362 & -0.0003 \\ -8.0348 & -14.4337 & 1.0566 \end{bmatrix}, \quad \hat{\Phi}_2 = \begin{bmatrix} 0.2831 & 0.0370 & 0.0003 \\ 0.2407 & -0.0928 & 0.0004 \\ 30.3958 & -12.9863 & -0.0821 \end{bmatrix}$$

$$\hat{\Phi}_3 = \begin{bmatrix} -0.0273 & 0.0999 & 0.0007 \\ -0.1673 & 0.0829 & 0.0020 \\ -14.3464 & 21.2713 & 0.2251 \end{bmatrix}, \quad \hat{\Phi}_4 = \begin{bmatrix} -0.0431 & 0.1041 & -0.0008 \\ -0.0766 & 0.1638 & -0.0019 \\ 8.1384 & -11.6987 & -0.3568 \end{bmatrix}$$

$$\hat{\Sigma} = \begin{bmatrix} 0.0018 & 0.0003 & -0.0176 \\ 0.0003 & 0.0050 & -0.1300 \\ -0.0176 & -0.1300 & 84.8262 \end{bmatrix}, \quad \hat{\alpha} = \begin{bmatrix} 0.0731 \\ 0.0509 \\ 9.9678 \end{bmatrix}$$

The corresponding model specification structure containing the standard errors of the ML parameter estimates (SE) is given below.

$$SE_{\hat{\Phi}_1} = \begin{bmatrix} 0.0881 & 0.0544 & 0.0004 \\ 0.1482 & 0.0916 & 0.0006 \\ 1.9363 & 1.1976 & 0.0833 \end{bmatrix}, \quad SE_{\hat{\Phi}_2} = \begin{bmatrix} 0.1026 & 0.0571 & 0.0006 \\ 0.1726 & 0.0960 & 0.0009 \\ 2.2561 & 1.2556 & 0.1236 \end{bmatrix}$$

$$SE_{\hat{\Phi}_3} = \begin{bmatrix} 0.1038 & 0.0574 & 0.0006 \\ 0.1747 & 0.0966 & 0.0009 \\ 2.2831 & 1.2630 & 0.1227 \end{bmatrix}, \quad SE_{\hat{\Phi}_4} = \begin{bmatrix} 0.0872 & 0.0550 & 0.0004 \\ 0.1468 & 0.0925 & 0.0006 \\ 1.91839 & 1.2093 & 0.0842 \end{bmatrix}$$

$$SE_{\hat{\Sigma}} = \begin{bmatrix} 0.0002 & 0.0003 & 0.0345 \\ 0.0003 & 0.0006 & 0.0592 \\ 0.0345 & 0.0592 & 10.7298 \end{bmatrix}, \quad SE_{\hat{\alpha}} = \begin{bmatrix} 0.0262 \\ 0.0440 \\ 3.7539 \end{bmatrix}$$

The results show clearly an one-way relationship between $X_t = (Y_1t, Y_2t)'$ variables and output Y_3t, since the AR coefficient estimates in the $\hat{\Phi}_j(1,3)$ and $\hat{\Phi}_j(2,3)$ positions are close to zero and relatively small to their perspective elements in $SE_{\hat{\Phi}_j}$. This highlights the exogeneity of X_t to Y_3t. In contrast, the higher values on the lower left corner of each of the $\hat{\Phi}_j$ matrices denote that only structured data variables Y_3t depend on the past values of semi-structured Y_1t and unstructured data Y_2t. Further, we can test our results by fitting a constrained AR(4) model, by conditional ML, so that $\hat{\Phi}_j(1,3) = 0$ and $\hat{\Phi}_j(2,3) = 0$. This leads to the following conditional ML estimates of $\hat{\Sigma}_0$:

$$\hat{\Sigma}_0 = \begin{bmatrix} 0.0018 & 0.0006 & -0.0234 \\ 0.0006 & 0.0051 & -0.1777 \\ -0.0234 & -0.1777 & 85.0146 \end{bmatrix}$$

with corresponding selection criteria, AIC_0 =-6.7046, $SBIC_0$ =-6.7101 and $HQIC_0$ =-6.9783. LR tests have been used to examine the hypothesis of the exogeneity constraints in the AR(4) model by comparing the fits between the unconstrained (full) and constrained model. This yields to $\acute{\lambda} = (mk + 3/2 - T + m) \log \frac{|\Sigma_m|}{|\Sigma_0|} = 4.5445$, whereas the constrained model has fewer parameters than the full model; in particular, there are eight degrees of freedom. Hence, it can be inferred that for this new method for testing it is not necessary to use the full multivariate model and the variables Y_1t and Y_2t can be considered as exogenous inputs to the output Y_3t. Furthermore, the distinct covariance variables $\hat{\Sigma}_0(1,3) = \hat{\Sigma}_0(3,1)$ and $\hat{\Sigma}_0(2,3) = \hat{\Sigma}_0(3,2)$ are sufficiently small, which indicate that m factors have accounted sufficiently well for the correlations of the variables and, thus, there is no need to include lag zero terms Y_1t and Y_2t towards Y_3t improvement.

Also, by examining the difference between the actual correlations and the modeled one, we found (relatively) very small values in all individual residual correlation matrices $\hat{r}_\varepsilon(\ell)$, which indicates a very good fit for the model. In particular, we obtained the following first four-ℓ (lags) residual correlation matrices $\hat{r}_\varepsilon(\ell) = \hat{V}D(\ell)\hat{V}$, where $D(\ell)$ is the covariance matrix at lag-ℓ and $\hat{V} = diag[d_{11,\ell}^{-1/2}, d_{22,\ell}^{-1/2}, d_{33,\ell}^{-1/2}]$, where $d_{ii,\ell}$ denotes the (i,i)th diagonal element of $D(\ell)$:

$$\hat{r}_\varepsilon(1) = \begin{bmatrix} 0.5764 & -0.1201 & -0.0002 \\ -0.0835 & 0.2362 & -0.0003 \\ -8.0348 & -14.4337 & 1.0566 \end{bmatrix}, \hat{r}_\varepsilon(2) = \begin{bmatrix} 0.2831 & 0.0370 & 0.0003 \\ 0.2407 & -0.0928 & 0.0004 \\ 30.3958 & -12.9863 & -0.0821 \end{bmatrix}$$

$$\hat{r}_\varepsilon(3) = \begin{bmatrix} -0.0069 & -0.0478 & -0.0042 \\ 0.0109 & -0.0210 & 0.0040 \\ 0.0570 & 0.0420 & -0.0131 \end{bmatrix}, \hat{r}_\varepsilon(4) = \begin{bmatrix} -0.1365 & 0.0271 & -0.0450 \\ 0.0724 & -0.0011 & 0.0151 \\ 0.0323 & 0.0775 & -0.1068 \end{bmatrix}$$

Note that the variables in residual correlation matrices $\hat{r}_\varepsilon(\ell)$ are very small, which indicate that adding an MA-term in our model would not lead to significant improvements over the from the fitted constrained AR(4) model. The final model based on the LS estimates is given below:

$$\hat{\Phi}_1 = \begin{bmatrix} 0.5862 & -0.1148 & 0 \\ -0.0579 & 0.2554 & 0 \\ -8.7538 & -14.9638 & 1.0475 \end{bmatrix}, \quad \hat{\Phi}_2 = \begin{bmatrix} 0.2767 & 0.0312 & 0 \\ 0.2251 & -0.0978 & 0 \\ 30.8355 & -12.8216 & -0.0690 \end{bmatrix}$$

$$\hat{\Phi}_3 = \begin{bmatrix} -0.0322 & 0.0854 & 0 \\ -0.1716 & 0.0571 & 0 \\ -14.2040 & 22.0240 & 0.2804 \end{bmatrix}, \quad \hat{\Phi}_4 = \begin{bmatrix} -0.0348 & 0.0768 & 0 \\ -0.0601 & 0.1063 & 0 \\ 7.6658 & -10.0511 & -0.4108 \end{bmatrix}$$

$$\hat{\alpha} = \begin{bmatrix} 0.0766 \\ 0.0638 \\ 9.6096 \end{bmatrix}$$

Last, in ML estimation methods we used for unconstrained and constrained models both corresponding Hessian matrices were positive-definite (all eigenvalues were positive). The constrained model is also stable since all eigenvalues of the associated lag operators have modulus less than 1 (Hamilton, 1994). In (5) we give the relationship between structured data Y_3t and semi-structured Y_1t and unstructured data Y_2t in the form of transfer function, where B is the lag operator.

$$\begin{aligned} &(1 - 1.0475B + 0.0691B^2 - 0.2804B^3 + 0.4109B^4)Y_3t = \\ &9.6097 - (14.9638B + 12.8216B^2 - 22.0240B^3 + 10.0511B^4)Y_{2,t-1} \\ &- (8.7539B - 30.8355B^2 + 14.2040B^3 - 7.6659B^4)Y_{1,t-1} + \varepsilon_3 t \end{aligned} \tag{5}$$

It can be inferred that the variables for structured data are mostly (however, slightly) influenced from changes of unstructured data. Thus, the constrained vector autoregressive model, in contrast to various univariate approaches, can not only capture the linear interdependencies among semi-structured and unstructured data time series but also provides additional identifying information that can be used to confront with the volatility of structured data. Also, since the VARX model involves current and lagged values of time series of different types of data, it may capture co-movements and underlying information that cannot be detected in univariate or bivariate models. This is an outcome, which is also highlighted in (Gutiérrez-Estrada et al., 2007; Cadenas et al., 2016; Diaz et al., 2016). Last, the transfer function, which is given in (5), provides a better understanding of the specific additional features of the multivariate model and may help in making more informed decisions how and what exogenous variables can be used in this particular approach.

5.1. Cointegration modelling

Next, we examine the non-stationarity features of time series data Y_t. The augmented Phillips-Perron tests based on autoregressive model with a drift variant for Y_t series, failed to reject the unit-root null hypothesis only for structured data Y_3t (Phillips and Perron, 1988, pp. 335-346). In order to further investigate the non-stationary nature of the structured data, we investigate the balance of relationships

in stable long-term levels on the variables by calculating the long-run AR matrix $\hat{\Pi} = -(I_3 - \Phi_1 - \Phi_2 - \Phi_3 - \Phi_4)$ in the final constrained AR(4) model, which is given by (6):

$$\hat{\Pi} = \begin{bmatrix} 0.2041 & -0.0786 & 0 \\ 0.0646 & 0.6789 & 0 \\ -15.5435 & 15.8124 & 0.1520 \end{bmatrix} \tag{6}$$

From (6), it can be inferred that the univariate model for Y_3t can be principally represented by an ARIMA (3,1,0) model by differencing $(1 - B)Y_t$. In fact, we can assert that Y_t contains a unit root and it is integrated to $I(1)$ if ΔY_t is stationary. The Johansen (1995) approach for cointegration allows for testing the validity of equilibrium relationships that govern the joint evolution of all three variables. We consider the following vector error-correction dynamic model by allowing the inclusion of a constant (trend-stationary) term δ (a k×1 vector of unknown constants) to capture the exogenous growth:

$$\Delta Y_t = \Pi Y_{t-1} + \sum_{j=1}^{p-1} \Gamma_j \Delta Y_{t-j} + \delta + \varepsilon_t \tag{7}$$

The LR tests reject the hypothesis for at least d=2 unit roots while fail to reject the hypothesis of the at least d=1 unit root ($r \leq 2$, where r= rank(Π)). The corresponding Gaussian estimator of $\tilde{\Pi} = AB'$, where A represents the speed of adjustment to equilibrium coefficients and B' represents the long run matrix of coefficient is given by (8) alongside with the corresponding residual covariance matrix $\tilde{\Sigma}$:

$$\tilde{\Pi} = \begin{bmatrix} 0.0317 & -0.1796 & 0 \\ -0.0922 & 0.5513 & 0 \\ -12.0283 & 19.1987 & 0.1558 \end{bmatrix} \quad \tilde{\Sigma} = \begin{bmatrix} 0.0018 & 0.0004 & -0.0205 \\ 0.0004 & 0.0050 & -0.1329 \\ -0.0205 & -0.1329 & 84.8926 \end{bmatrix} \tag{8}$$

Note that the eigenvalues of $\tilde{\Pi}$ are -0.5131, 0.0000 and -0.1651. Hence, we consider a reduced-rank estimation of Π with rank(Π)=2, which incorporates the unit root while testing $H_2 : rank(\Pi) = 2$. In (8), the single cointegrated linearly relationships that are indicated in the partial canonical correlation analysis associated with the LR testing is found to be $Z_3t = -12.0283Y_2t + 19.1987Y_3t + 0.1558Y_1t$, which is similar to the linear combination that is derived from the third row of the estimated long-run matrix for the final constrained AR(4) model given in (6). The Z_3t is clearly an $I(0)$ and, therefore, represents the stable long-term relationship between structured, semi-structured and unstructured data. Figure 4 shows the linear nature of the transformed variables Z_3t, which follow a clear stationary trend and signalises the stable long-term relationship between semi-structured and unstructured data.

5.2. Performance analysis and forecasting

In this section, we compare the performance of the constrained VARX model in terms of estimation and predicative accuracy with other three linear time series models, with techniques that are based on structured data and used broadly in literature. In

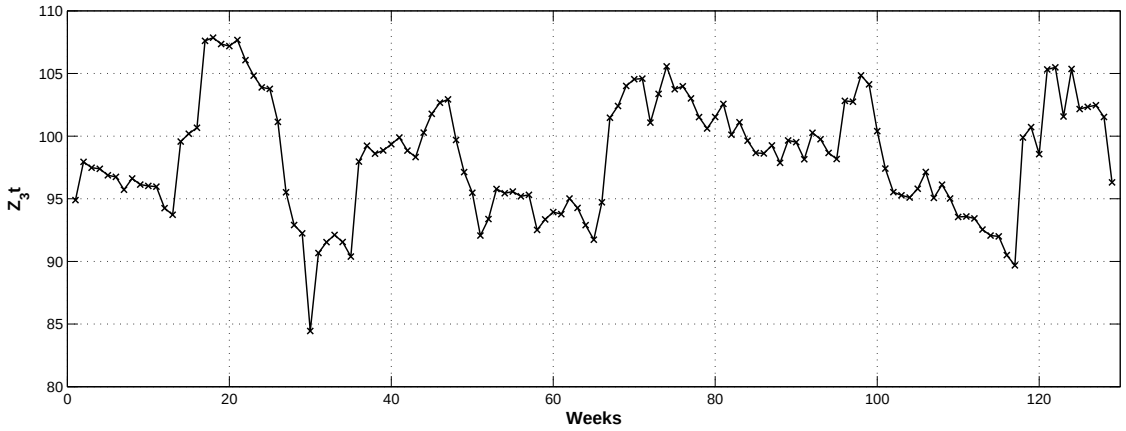

Figure 4: Cointegrated time series $Z_3t = -12.0283Y_2t + 19.1987Y_3t + 0.1558Y_1t$

particular, we consider the following ML estimation of a non-seasonal ARIMA(4,1,4) model; *Model1* as: (1 - 0.5020B - 0.6182B^2-0.5735B^3+0.8639B^4)(1 - B)Y_3t = + 0.0166 (1 - 0.3594B -0.7259B^2 -0.6104B^3+0.6958B^4)ε_t.

In the second model we include exogenous semi-structured and unstructured data $X_t = (Y_1t, Y_2t)'$, and consider an ARIMAX(4,1,4) model, *Model2* as: (1 - 0.5103B - 0.4167B^2 -0.4528B^3 -0.5958B^4)(1 - B^{-1})Y_3t = -0.0600 + [2.0501 -13.7024]X_t + (1 + 0.3797B + 0.5413B^2 +0.3762B^3 -0.2973B^4)ε_t.

Last, we specify a multiplicative seasonal ARIMA$(0,1,1)(0,1,1)_{52}$ model with seasonal and nonseasonal integration in order to capture any trend and seasonality in Y_3t time series, *Model3* as follows: (1 - B)(1 - B^{52})Y_3t = 0.0759 + (1 + 0.0785)(1 + 0.5963B^{52})ε_t

Table 3 summarises the main results of three univariate models against AIC, BIC criteria for estimation (Makridakis et al. (2008)). Note, that we have also considered in our analysis the specification of an additive seasonal ARIMA(0,1,1) at lag 52, however, its performance in terms of estimation and forecasting did not provide better results in comparison with the multiplicative one (*Model3*).

Table 3: Performance measures for all three models

Model	**AIC**[1]	**BIC**[2]	**MSE**[3]	**MAE**[3]
Model1	800.8584	828.4619	1156.7442	30.9761
Model2	800.0991	834.8938	457.5686	19.6883
Model3	897.0827	908.4278	1609.9878	34.1487
VARX	-	-	430.4882	13.3893

[1] $AIC = log\hat{\sigma^2} + 2(p+q)/N$,

[2] $BIC = log(\hat{\sigma^2}) + 2(p+q)log(N)/N$,

[3] Computation in VARX only for Y_3t

Figure 6 illustrates the performance of all univariate models in terms of forecasting.

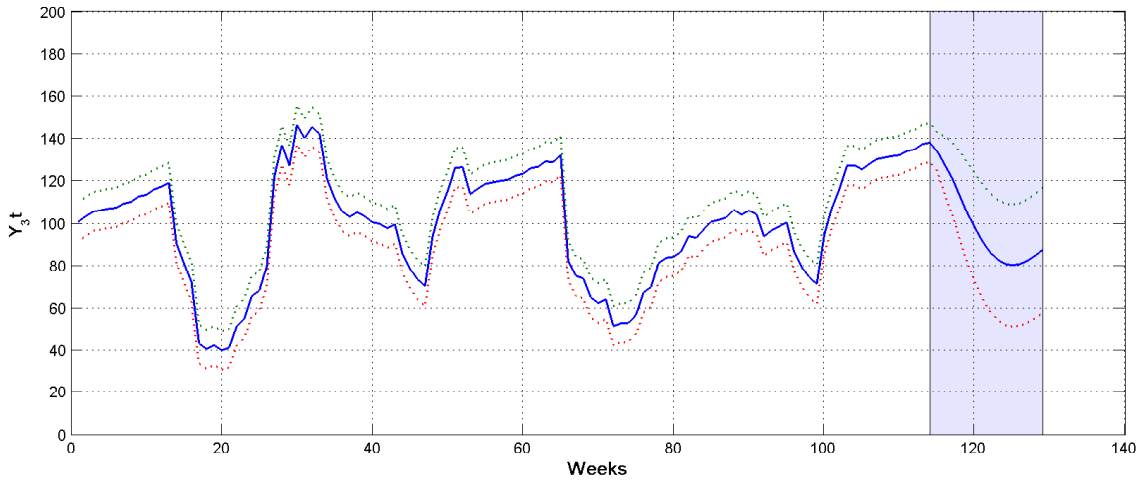

Figure 5: Forecast of series Y_3t for VARX model with (1-σ) lower and upper limits

It is evident that the inclusion of the linear effect of exogenous covariates in *Model2* provides better results against the simple ARIMA model *Model1* and seasonal-based *Model3*. Figure 5 depicts the forecasting of Y_3t time series based on the constrained VARX(4) model. It can be easily inferred from the stationary response series Y_3t that multivariate model slightly outperforms univariate models by including both exogenous series in the 3-dimensional multivariate time series. This is also supported by the results shown in Table 3, where the VARX(4) model presents the smallest mean-square error (MSE) and mean absolute error (MAE) for Y_3t time-series in comparison with all three univariate models.

Note, that the models proposed in this paper involve linear components of big data series that the univariate models might not be able to extract. Also, our model, under relaxing stationarity assumptions, is reinforced by the fact that multiple data series are triggered by common sources (patients) and this helps to generalise univariate autoregressive models by allowing for more than one evolving variable. We understand that there is a lot of debate about the performance of multivariate models against various univariate approaches in terms of forecasting when volatility is present on demand patterns. However, in this paper we highlight the advantages of inclusion of relevant information from underlying different data structures from mainly common sources as it also suggested by (Lütkepohl and Netšunajev, 2017; Stock and Watson, 2001).

6. Conclusions

The challenge and impact of demand volatility and forecast inaccuracy experienced by retail pharmacies have driven the need for more advanced quantitative techniques, which consider big data from exogenous sources that indicate consumer behaviour and have emerged only in recent years. This research was embarked upon to contribute to this existing gap. Our results support the assumption that through the use of multivariate time series analysis technique (VARX), customer generated content from a variety of internet sources (Google, YouTube and Newspapers) has the potential

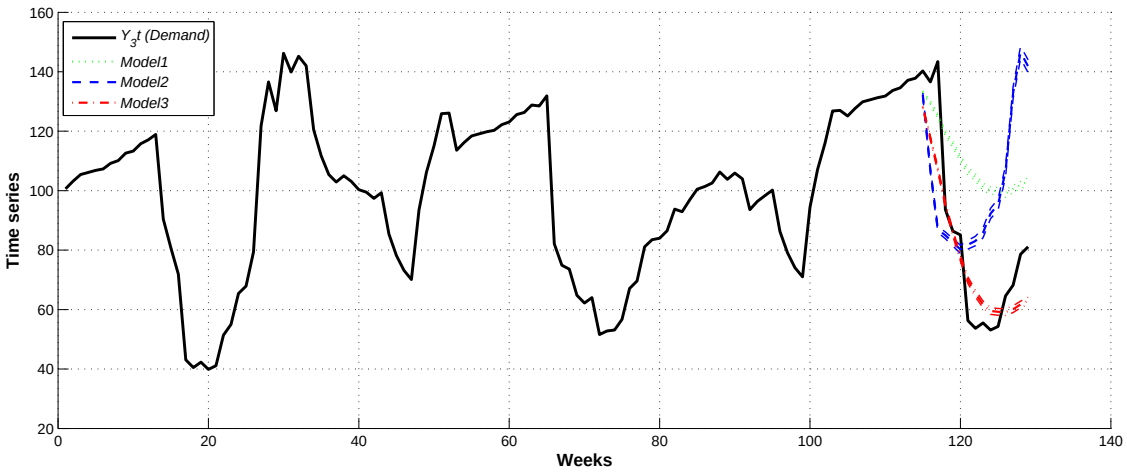

Figure 6: Forecast of series Y_3t for different univariate models

of reducing demand forecasting errors in the short term and improving response to demand volatility in retail pharmacy settings.

This paper highlights also the importance of business intelligence through data extraction, mining and analytics to forecasting and demand planning. As discussed earlier, text mining and analytics are applied with great success in a variety of industries, especially in the retail sector. However, there is little evidence of the application of this data-driven approach in retail pharmacy demand planning. Considering the results from Anusha et al. (2014) and Khalil Zadeh et al. (2014), which have highlighted the immense consequences of inaccurate forecasting in retail pharmacy settings, it is believed that our results will encourage further investigation in this direction thereby proffering innovative solutions to tackle this challenge. Furthermore, with the increasing government regulation in the sale of medicines and fierce competition by rivals, which continue to reduce business profits, the incorporation of exogenous user generated content into multivariate forecasting techniques like VARX could provide the needed competitive advantage for retail pharmacy companies.

The proposed approach in this paper is a good choice to be used as a comparator to other estimation and forecasting techniques in practice. Although results from this study elucidate the importance of quantitative techniques, it is important to state that better results may be achieved by additionally considering other internal and external factors occurring in practice and which could impact the performance of our model. Although this research attempted to evaluate the impact of the five big data characteristics, it can be argued that it majorly focused on variety through the consideration of data from different sources. However, the encouraging results indicate that the impact of other big data characteristics (velocity, volume, value, veracity) can also be evaluated. As 129 weeks (two and half years) volume of data was considered, it is possible to scale up the volume depending on data availability.

Methodologically, it can be argued that this research has innovatively pioneered the use of a combination of consumer behaviours such as Google search intensity, YouTube video watching and newspaper reading as key variables in predicting demand

for medicines. Although Google search intensity had been used in the past for disease outbreak prediction (influenza), there is a lack of evidence showing its use in medicine demand prediction. The same also goes for YouTube and newspaper reading. As we also observed that the increase in the demand of OTC analgesics coincided with the African Ebola outbreak in 2014, it may be noteworthy to consider if the Zika virus epidemic, which has fever as a symptom and broke out in the Americas in 2015, has the same effect on related OTC medicines.

Empirically, we believe that having utilised real life data to tackle a real challenge of a real life company and achieved realistic and acceptable results, this represents a significant step in the use of big data analytics in pharmaceutical retail demand planning. Theoretically, it is hoped that based on our findings, this study has contributed to much-needed knowledge lacking in this area and this would encourage the development of new types of demand forecasting models especially those considering customer generated content in pharmaceutical retail demand planning. Last, considering that demand volatility is a challenge across retail industries, it may be of note to consider if the results from this study would be consistent if replicated in related sectors like food or apparel industries.

References

Anusha, S.L., Alok, S., Ashiff, S., 2014. Demand forecasting for the Indian pharmaceutical retail: A case study. Journal of Supply Chain Management Systems 3.

Askitas, N., Zimmermann, K.F., 2009. Google econometrics and unemployment forecasting. Applied Economics Quarterly 55, 107–120.

Azadeh, S.S., Marcotte, P., Savard, G., 2015. A non-parametric approach to demand forecasting in revenue management. Computers & Operations Research 63, 23–31.

Bansal, K., Vadhavkar, S., Gupta, A., 1998. Neural networks based data mining applications for medical inventory problems. International Journal of Agile Manufacturing 1, 187–200.

Barfar, A., Padmanabhan, B., 2015. Does television viewership predict presidential election outcomes? Big data 3, 138–147.

Betts, J.M., 2014. Calculating target inventory levels for constrained production: A fast simulation-based approximation. Computers & Operations Research 49, 18–27.

Borghol, Y., Mitra, S., Ardon, S., Carlsson, N., Eager, D., Mahanti, A., 2011. Characterizing and modelling popularity of user-generated videos. Performance Evaluation 68, 1037–1055.

Box, G.E., Jenkins, G.M., Reinsel, G.C., Ljung, G.M., 2015. Time series analysis: forecasting and control. John Wiley & Sons.

Bratina, D., Faganel, A., 2008. Forecasting the primary demand for a beer brand using time series analysis. Organizacija 41, 116–124.

Buzia, O.D., Mardare, N., Diaconu, C., 2016. Forecast of pharmacy sales with Brown's exponential smoothing. Acta Medica Transilvanica 21.

Cadeaux, J., Dubelaar, C., 2012. Market environment, assortment policy, and performance of small retailers. Australasian Marketing Journal (AMJ) 20, 250–259.

Cadenas, E., Rivera, W., Campos-Amezcua, R., Heard, C., 2016. Wind speed prediction using a univariate ARIMA model and a multivariate NARX model. Energies 9, 109.

Canadian consumer health products, 2012. Over-The-Counter medications, viewed on June 2016, http://www.chpcanada.ca/en/industry-products/faqs/over-counter-medications.

Carrizosa, E., Olivares-Nadal, A.V., Ramírez-Cobo, P., 2016. Robust newsvendor problem with autoregressive demand. Computers & Operations Research 68, 123–133.

Chao, L., 2015. Big data brings relief to allergy medicine supply chains, the Wall Street journal, viewed on 17 August 2016, http://www.wsj.com/articles/big-data-brings-relief-to-allergy-medicine-supply-chains-1432679948.

Chen, H., Chiang, R.H., Storey, V.C., 2012. Business intelligence and analytics: From big data to big impact. MIS quarterly 36, 1165–1188.

Choi, H., Varian, H., 2012. Predicting the present with Google trends. Economic Record 88, 2–9.

Dean, J., Ghemawat, S., 2008. Mapreduce: simplified data processing on large clusters. Communications of the ACM 51, 107–113.

Diaz, E.M., Molero, J.C., de Gracia, F.P., 2016. Oil price volatility and stock returns in the G7 economies. Energy Economics 54, 417 – 430.

Ericsson consumer lab, 2015. Internet Goes Mobile: Country report Nigeria. an Ericsson consumer insight summary report over-the-counter medications, viewed on 22 July 2016, https://www.ericsson.com/res/docs/2015/consumerlab/ericsson-consumerlab-internet-goes-mobile-nigeria.pdf.

Fahimnia, B., Davarzani, H., Eshragh, A., 2015. Planning of complex supply chains: A performance comparison of three meta-heuristic algorithms. Computers & Operations Research .

Ferreira, K.J., Lee, B.H.A., Simchi-Levi, D., 2015. Analytics for an online retailer: Demand forecasting and price optimization. Manufacturing & Service Operations Management 18, 69–88.

Figueiredo, F., 2013. On the prediction of popularity of trends and hits for user generated videos, in: Proceedings of the sixth ACM international conference on Web search and data mining, ACM. pp. 741–746.

Fiore, S., D'Anca, A., Palazzo, C., Foster, I., Williams, D.N., Aloisio, G., 2013. Ophidia: toward big data analytics for escience. Procedia Computer Science 18, 2376–2385.

Gandomi, A., Haider, M., 2015. Beyond the hype: Big data concepts, methods, and analytics. International Journal of Information Management 35, 137–144.

Gibson, S., 2012. Direct-to-consumer advertising in the digital age: The impact of the internet and social media in the promotion of prescription drugs in canada. Ph.D. thesis. University of Toronto.

Greene, J.A., Kesselheim, A.S., 2010. Pharmaceutical marketing and the new social media. New England Journal of Medicine 363, 2087–2089.

Gutiérrez-Estrada, J.C., Silva, C., Yáñez, E., Rodríguez, N., Pulido-Calvo, I., 2007. Monthly catch forecasting of anchovy Engraulis ringens in the north area of Chile: non-linear univariate approach. Fisheries Research 86, 188–200.

Hamilton, J.D., 1994. Time series analysis. Princeton University Press, NJ, USA.

Hamuro, Y., Katoh, N., Matsuda, Y., Yada, K., 1998. Mining pharmacy data helps to make profits. Data Mining and Knowledge Discovery 2, 391–398.

Hazen, B.T., Boone, C.A., Ezell, J.D., Jones-Farmer, L.A., 2014. Data quality for data science, predictive analytics, and big data in supply chain management: An introduction to the problem and suggestions for research and applications. International Journal of Production Economics 154, 72–80.

Jain, A., Menon, M.N., Chandra, S., 2014. Sales Forecasting for Retail Chains. Working Paper.

Johansen, S., 1995. Likelihood-based inference in cointegrated vector autoregressive models. Oxford University Press on Demand.

Jones, E.C., Gupta, S., 2015. Hospital supply chain management by implementing RFID. International Journal of Supply Chain Management 4.

Jones, S.S., Evans, R.S., Allen, T.L., Thomas, A., Haug, P.J., Welch, S.J., Snow, G.L., 2009. A multivariate time series approach to modeling and forecasting demand in the emergency department. Journal of biomedical informatics 42, 123–139.

Kenny, D.A., 2011. Terminology and basics of SEM, viewed on May 2017, http://davidakenny.net/cm/basics.htm.

Khalil Zadeh, N., Sepehri, M.M., Farvaresh, H., 2014. Intelligent sales prediction for pharmaceutical distribution companies: A data mining based approach. Mathematical Problems in Engineering 2014.

Kholodilin, K.A., Thomas, T., Ulbricht, D., 2014. Do media data help to predict German industrial production? DIW Berlin Discussion Paper.

Kim, W., Won, J.H., Park, S., Kang, J., 2015. Demand forecasting models for medicines through wireless sensor networks data and topic trend analysis. International Journal of Distributed Sensor Networks 2015, 36.

Kirian, M.L., Weintraub, J.M., 2010. Prediction of gastrointestinal disease with over-the-counter diarrheal remedy sales records in the San Francisco Bay area. BMC medical informatics and decision making 10, 1.

Laney, D., 2001. 3-D data management: Controlling data volume, velocity and variety. META Group Inc. Original research note.

Lee, H.L., Padmanabhan, V., Whang, S., 1997. Information distortion in a supply chain: The bullwhip effect. Management science 43, 546–558.

Lee, J., Lapira, E., Bagheri, B., Kao, H.a., 2013. Recent advances and trends in predictive manufacturing systems in big data environment. Manufacturing Letters 1, 38–41.

Li, X., Ma, J., Wang, S., Zhang, X., 2015. How does Google search affect trader positions and crude oil prices? Economic Modelling 49, 162–171.

Linoff, G.S., Berry, M.J., 2011. Data mining techniques: for marketing, sales, and customer relationship management. John Wiley & Sons.

Liu, Q., Zhang, X., Liu, Y., Lin, L., 2013. Spreadsheet inventory simulation and optimization models and their application in a national pharmacy chain. INFORMS Transactions on Education 14, 13–25.

Lütkepohl, H., Netšunajev, A., 2017. Structural vector autoregressions with heteroskedasticity: A review of different volatility models. Econometrics and Statistics 1, 2–18.

Mahar, S., Salzarulo, P.A., Wright, P.D., 2012. Using online pickup site inclusion policies to manage demand in retail/e-tail organizations. Computers & Operations Research 39, 991–999.

Makridakis, S., Wheelwright, S.C., Hyndman, R.J., 2008. Forecasting methods and applications. John Wiley & Sons.

McNulty, E., 2014. Understanding big data: The seven Vs. http://dataconomy.com/seven-vs-big-data/ (viewed on 12 June 2016) .

Nardo, M., Petracco, M., Naltsidis, M., 2016. Walking down wall street with a tablet: A survey of stock market predictions using the web. Journal of Economic Surveys 30, 356–369.

Obu, H.A., Chinawa, J.M., Ubesie, A.C., Eke, C.B., Ndu, I.K., 2012. Paracetamol use (and/or misuse) in children in Enugu, South-East, Nigeria. BMC pediatrics 12, 1.

Ohlhorst, F.J., 2012. Big data analytics: turning big data into big money. John Wiley & Sons.

Pan, B., Chenguang Wu, D., Song, H., 2012. Forecasting hotel room demand using search engine data. Journal of Hospitality and Tourism Technology 3, 196–210.

Papanagnou, C., Halikias, G., 2008. Supply-chain modelling and control under proportional inventory-replenishment policies. International Journal of Systems Science 39, 699–711.

Phillips, P.C., Perron, P., 1988. Testing for a unit root in time series regression. Biometrika 75, 335–346.

Power, D.J., 2014. Using big data for analytics and decision support. Journal of Decision Systems 23, 222–228.

Reinsel, G.C., 2003. Elements of multivariate time series analysis. Springer Science & Business Media.

Riahi, N., Hosseini-Motlagh, S.M., Teimourpour, B., 2013. A three-phase hybrid times series modeling framework for improved hospital inventory demand forecast. International Journal of Hospital Research 2, 133–142.

Ribeiro, A., Seruca, I., Durỳo, N., 2016. Sales prediction for a pharmaceutical distribution company: A data mining based approach, in: Information Systems and Technologies (CISTI), 2016 11th Iberian Conference on, AISTI. pp. 1–7.

Sawon, M., Hasan, T., Hosen, M., et al., 2016. Prediction on large scale data using extreme gradient boosting. Ph.D. thesis. BRAC University.

Shankar, V., Li, J.K., 2014. Leveraging social media in the pharmaceutical industry, in: Innovation and marketing in the pharmaceutical industry. Springer, New York, NY, pp. 477–505.

Socan, M., Erculj, V., Lajovic, J., 2012. Early detection of influenza like illness through medication sales. Central European journal of public health 20, 156.

Stock, J.H., Watson, M.W., 2001. Vector autoregressions. The Journal of Economic Perspectives 15, 101–115.

Sun, D., Du, Y., Xu, W., Zuo, M., Zhang, C., Zhou, J., 2015. Combining online news articles and web search to predict the fluctuation of real estate market in big data context. Pacific Asia Journal of the Association for Information Systems 6.

Suziedelyte, A., 2012. How does searching for health information on the Internet affect individuals' demand for health care services? Social science & medicine 75, 1828–1835.

Thomassey, S., Fiordaliso, A., 2006. A hybrid sales forecasting system based on clustering and decision trees. Decision Support Systems 42, 408–421.

Tuhkuri, J., et al., 2016. ETLAnow: A Model for Forecasting with Big Data Forecasting Unemployment with Google Searches in Europe. Technical Report. The Research Institute of the Finnish Economy.

Tyrawski, J., DeAndrea, D.C., 2015. Pharmaceutical companies and their drugs on social media: a content analysis of drug information on popular social media sites. Journal of medical Internet research 17, e130.

Verbeke, W., Ward, R.W., 2001. A fresh meat almost ideal demand system incorporating negative TV press and advertising impact. Agricultural Economics 25, 359–374.

Waller, M.A., Fawcett, S.E., 2013. Data science, predictive analytics, and big data: a revolution that will transform supply chain design and management. Journal of Business Logistics 34, 77–84.

Wang, G., Gunasekaran, A., Ngai, E.W., Papadopoulos, T., 2016. Big data analytics in logistics and supply chain management: Certain investigations for research and applications. International Journal of Production Economics 176, 98–110.

Watson, J.W., Moliver, N., Gossett, K., 2014. Inventory control methods in a long-term care pharmacy: Comparisons and time-series analyses. Journal of Pharmacy Technology 30, 151–158.

Wosinska, M., 2002. Just what the patient ordered? direct-to-consumer advertising and the demand for pharmaceutical products. HBS Marketing Research Paper .

Yadav, P., 2015. Health product supply chains in developing countries: Diagnosis of the root causes of underperformance and an agenda for reform. Health Systems & Reform 1, 142–154.

Zhang, X., Meiser, D., Liu, Y., Bonner, B., Lin, L., 2014. Kroger uses simulation-optimization to improve pharmacy inventory management. Interfaces 44, 70–84.