Ann Oper Res (2012) 192:105–121
DOI 10.1007/s10479-011-0904-5

A data mining framework for product and service migration analysis

Siu-Tong Au · Rong Duan · Wei Jiang

Published online: 21 May 2011

Abstract With new technologies or products invented, customers migrate from a legacy product to a new product from time to time. This paper discusses a time series data mining framework for product and service migration analysis. In order to identify who migrate, how migrations look like, and the relationship between the legacy product and the new product, we first discuss certain characteristics of customer spending data associated with product migration. By exploring interesting patterns and defining a number of features that capture the associations between the spending time series, we develop a co-integration-based classifier to identify customers associated with migration and summarize their time series patterns before, during and after the migration. Customers can then be scored based on the migration index that integrates the statistical significance and business impact of migration customers. We illustrate the research through a case study of internet protocol (IP) migration in telecommunications and compare it with likelihood-ratio-based tests for change point detections.

Keywords Co-integration · Customer relationship management · Internet Protocol (IP) · Likelihood ratio tests · Telecommunication · Virtual Private Network (VPN)

1 Introduction

Michael Porter (1979) postulates five basic competitive forces for modern enterprizes, the treat of new entrances into the industry, the bargaining power of suppliers to the industry, the bargaining power of customers or buyers, the treat of substitute products or services, and the rivalry among existing firms. The rapid growth of new technologies nowadays makes the

S.-T. Au · R. Duan
AT&T Research Labs, Florham Park, NJ 07932, USA

W. Jiang (✉)
Antai College of Economics and Management, Shanghai Jiaotong University, Shanghai, P.R. China
e-mail: jiangwei08@gmail.com

industries not only face with competition in the market place, but also competition from the technology itself. Different from the market competition which is typically seen as a price challenge and services diversity from different competitors, technology competition comes in many forms in various industries, the most basic being the evolution of advanced services. As a consequence, market competition and technology competition influence each other and make enterprize decisions even harder.

As new technologies, products, and services penetrate to our daily life, it is possible to have billions of dollars revenue impact for a single company involves into migration of legacy products to new products. The next section provides many business examples of product migration. To gain as much value as possible from the existing legacy products and at the same time allocate sufficient resources to migrate legacy products to new products in order to keep competitiveness are ultimate projects for every industry. For example, the telecommunication industry is recently undergoing serious changes as people are increasingly turning to new technologies such as internet protocol (IP) and wireless as their primary mode of communication. It is therefore crucial for companies to understand the revenue gained from new products related to the lost of legacy products, which will assist to make decisions on when, how and who should be encouraged in product migration.

Business entities are frequently interested in the following questions related to product/service migration:

$Q1$: *What kind of information should be used to identify migration? How to define migration that makes strong business sense and catches business concerns?*
$Q2$: *How is the decline in the legacy product related to the growth in the new product? How much of the decline is due to migration or losses?*
$Q3$: *How does the migration look like? Do we lose any revenue due to migration and gain any afterward?*
$Q4$: *When and at what level will the legacy and new products stabilize?*
$Q5$: *What percentage of the legacy product has already migrated?*
$Q6$: *Who are migrating from the legacy to the new product?*
$Q7$: *Which customers are at risk and should be migrated?*

$Q1$ is the fundamental question in all product and service migration studies. It is usually assumed known or given through product/service ordering systems or other operational systems. However, as discussed later, it is not obvious to many corporations with complex enterprize structures and bureaucracies and needs to be clarified before any migration analysis. $Q2$–$Q4$ arise from product view and associate the overall revenue trend to the impacts imposed by product/service migrations. They are related to important corporation strategies on how and when to launch new products. On the other hand, $Q5$–$Q7$ tackle customer level analysis in operations and relate to the key issues of customer relationship management (CRM) on how to quantify and predict the migration activities at customer levels. While $Q2$–$Q3$ and $Q5$–$Q6$ look backwards to explore customer behaviors, $Q4$ and $Q7$ are forward looking and focus on finding explanatory variables that have the predictive power for migration customers. Once potential migration customers are identified, it is possible to predict the overall revenue trend for enterprizes so that resources can be allocated accordingly to satisfy customer demands. Nevertheless, $Q4$ and $Q7$ are not discussed here. Interested readers may refer to Murynets et al. (2009) for discussions of related research.

The objective of this paper is to propose a data mining framework for product migration analysis from enterprize recorded revenue systems. The data mining framework, stemmed on business concerns, can help enterprizes to explore their revenue history and identify patterns that drive the declines of the legacy product and/or the increases of the new product. In

order to flag migration customers out of millions of customer records, we construct a migration index and a set of features to quantify customer migrations, and develop a classifier to identify the customers who have migrated and who are migrating during a certain period of time. For simplicity, we focus our research on two products (services, programs, locations, etc.) only—one legacy product which typically has overall decline revenue history and the other new product which typically has overall increasing revenue history.

The rest of the paper is organized as follows. Section 2 discusses the business impact of product migration and reviews the current research related to product migration. Section 3 introduces the data mining framework that defines, identifies and quantifies the migration problem. In particular, a co-integration model is introduced to summarize migration patterns and a migration index is defined to integrate the statistical significance and business impacts of migration customers. Section 4 illustrates the implementation of our framework using a case study of IP migrations and compares it with the traditional change point detection method. Section 5 concludes the paper with ongoing research directions. The Appendix provides an overview of methods for co-integration analysis.

2 Product migration analysis

In order to develop a data mining framework for migration analysis, we first discuss the business impacts of product migrations and relevant research for migration identifications in marketing science and engineering.

2.1 Business impacts of product migration

Product migration, which is also called product substitution, refers to the replacement between two products due to product differentiation. Product migration has a profound impact on nowadays business due to intensive competitions, both internally and externally. Take the telecommunication industry as an example, it is fast-paced and constantly innovating, where technology substitution is a common occurrence and a serious concern for service providers. With the arrival of IP and multi-protocol label switching (MPLS) technologies, the old data networks are rapidly being replaced by next generation networks and IP migration is becoming a business imperative. According to an analysis by Yankee Group (2007), the global corporate data service packet revenue was declined from 24 to 19 billion dollars from 2003 to 2007 and is forecasted to further decline to 11 billions in 2011. On the other hand, IP VPN revenue was increased from 8 to 25 billions during the same period and is forecasted to increase to 44 billions in 2011. From the consumer side, the three Reginal Bell Operating Companies (RBOC), AT&T, Verizon and Qwest, have lost 17.2 million residential telephone lines and billions of dollars in revenue since the beginning of 2005. On the other hand, VoIP service providers have gained 14.4 million customers (TeleGeography 2008) at the same time.

Similar situations are happening in other competitive industries as well. The development of electronic information technologies and the spread of Internet lead a wide variety of E-commerce growth extraordinary and a sustained double-digit year-over-year growth rate. For example, in retail industry, the constant growing of online retailers is a big challenge to traditional retailers. According to Forrest Group, US online retail sales reached 175 billion dollars in 2007 and is projected to grow to 335 billions by 2012. The well-known DVD rental providers Blockbuster lost 1.2 billion dollars in 2004 and over 500 million (Andrew 2007) in 2005 due to the new technology based online service providers such as Netflix,

Apple, etc. Other DVD provider companies such as Hollywood Video even filled bankruptcy in 2007 (Reisinger 2008). Pharmaceutical industry is another innovation prone industry, whose performance heavily depends on companies' R&D pipelines. When a new drug or a second generation drug hits the market, the revenue for the old drug will gradually shift to the new drug, or lose to competitors. It is important for the company to know how much new drug revenue is from new customers, and how much are transferred from old drug. For example, Merck's Zocor, which was No. 2 cholesterol drug in the 15-billion-dollar US market, lost US patent protection in the mid 2006. Zocor sales fell 65%, partially due to other cheaper generic copies after the patent expired. At the same time, Merck's newer cholesterol pills, Vytorin and Zetia, rose 46% to 1.1 billion dollars (Bloomberg 2007). It is therefore important for Merck to understand how much Vytorin's gain is from Zocor's lose.

2.2 The state-of-the-art research of product migration

Product life cycle modeling lays down the groundwork for many product migration analysis in consumer research. The well-know S-shaped curve for product life cycle explains different stages of product development in market. Mahajan et al. (1990) provides an excellent review of the S-shaped curve and extensions of the Base model (Base 1969) to analyze the process of product substitutions.

In order to make prompt reactions to market changes, it is far from enough to conclude the product migration problem in the overall enterprize view only. To address the aforementioned questions, it is important to study the migration patterns for individual customers who purchase the products. Currently the major approach for migration analysis is through surveys. Roger (1976) suggests surveying customers at random to study their personal, demographical characteristics and communication behaviors, and find out the underlying migration reasons. Recently, Constantiou, et al. (2008) survey the Denmark IP telephony and identify the key economic factors and their impacts on the diffusion process. IBM (2008) surveys 600 consumers to study the substitution of Internet on PC with Internet on mobile. Allenet et al. (2003) study the substitution of branded drugs by generic drugs through survey of French market. Besides the survey analysis, Johnson and Bhatia (1997) propose a regression model to analyze the process of substitution on the land mobile radio market; Meuter et al. (2005) use multiple and logistic regression models to analyze self-service and clerk-based service substitutions; Prins (2008) studies the adoption of mobile phone use split-harzard approach.

Most of these studies focus on finding the explanatory variables for adoption likelihood and most importantly assume that the migration customers have been correctly labeled. There is few discussions on how to identify the migration customers besides surveys. It is apparent that surveys are often with limited scope and its accuracy is affected by survey designs such as sample size, representability of samples and other factors, and therefore questionable in certain level of confidence. In addition, complex structures and bureaucracies of corporate accounting systems also bring critical challenges to the survey analysis as "customers" themselves may not be well defined. It is the objective of this paper to develop a framework for migration analysis in the customer level that can help explain the driving forces of corporate revenue and lift the profit margins of marketing campaigns.

3 A data mining framework for customer migration analysis

In this section, we will develop a framework for migration analysis using enterprize billing systems. First we will discuss why the billing data is used in our analysis, followed by

an elaboration of important features presented in the customer billing data. Based on the customer features, we propose a co-integration model for billing time series of legacy and new products and define a migration indicator which integrates statistical significance and business impacts of migration customers.

3.1 What data should be used for migration analysis

As discussed in the last section, most existing research on customer level migration analysis assume that the migration status is known and customer activity information can be obtained from either ordering systems or customer surveys. It is assumed that ordering systems should capture all customer status changes such as the time for customer activation, inactivation, upgrade, migrate and other contract information. While this may be valid for consumer products with simple accounting structures, it is somewhat difficult to label business customers in some service companies where subdivisions provide different services and usually enjoy their own jurisdictions that prevent a comprehensive view of most enterprize customers. In such complex organizations, it is a big challenge to trace customer activities and streamline multiple billing accounts into a single customer view in order to fully understand the customers, especially when the enterprize has gone through business changes such as merging, acquisitions, spin-offs, etc. More importantly, the account activities recorded in the ordering system can not accurately provide an estimate of the revenue impact, which is mostly concerned for service/product providers.

On the other hand, many corporations record customer level spending and usage (voice and data services) by products overtime in the billing systems. Billing may not be as flashy a business function as CRM or sales, but it is almost unquestioned as a key strategic driver for companies in search of high performance. It reveals the revenue portfolio of the entire company and plays an important role in any revenue related marketing campaigns. There is considerable behavior information contained in thousands or millions of customer billing time series. Exploring patterns of these historical billing information provides an opportunity to investigate product migrations through data mining techniques.

3.2 Data explorations for customer migration analysis

While data exploration plays an important role in the knowledge discovery and data mining process, efforts on pertinent aspects of data are necessary for building meaningful models, especially for revenue time series data. In this migration study, based on data exploration and business knowledge, customers are classified into four groups by their business characteristics—new, disconnect, migration, and others. Figure 1 presents revenue time series of the legacy and new products for several typical customers in a telecommunication company. In the graph, the letter "L" stands for the legacy product and "N" stands for the new product. Due to proprietary reasons, the scale of time series have been removed and only the trend patterns are apparent from the graphs.

(a) *New customers* Figure 1(a) presents the revenue time series of a new customer whose legacy product revenue doesn't change during the recorded history but with increasing revenue for the new product from certain points within the time frame. The customer may have no legacy product with the company before or have been using the legacy product for a while and continuously generating revenue from it. The revenue time series patterns indicate a totally new customer to the company or an existing customer just adopting the new technology in service. New customers may be wins from competitors and are the type of customers that the corporation likes the most.

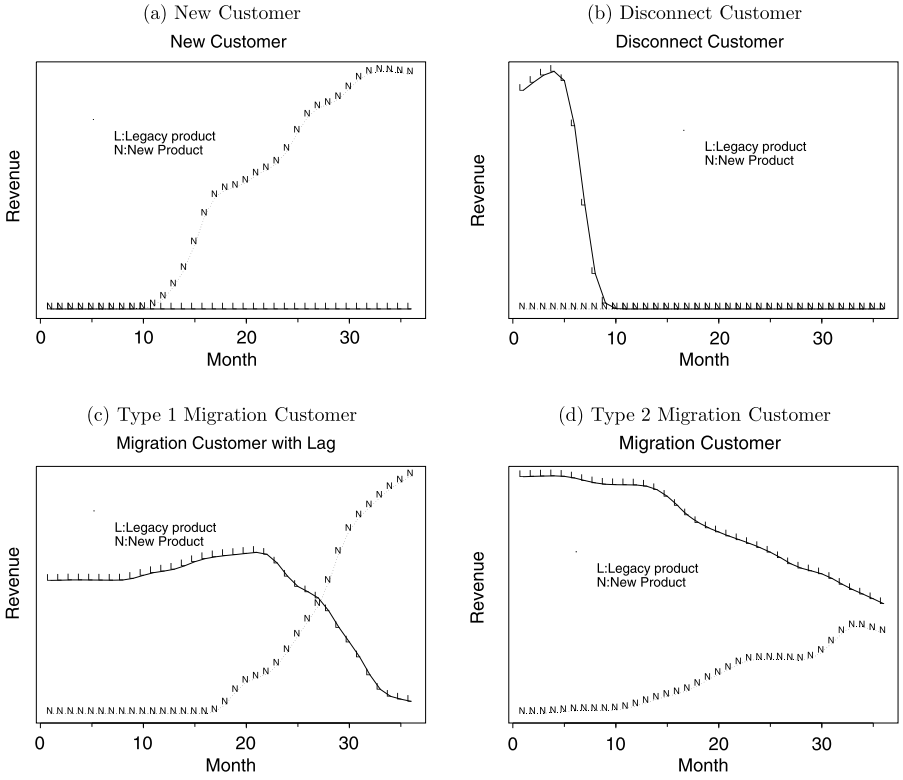


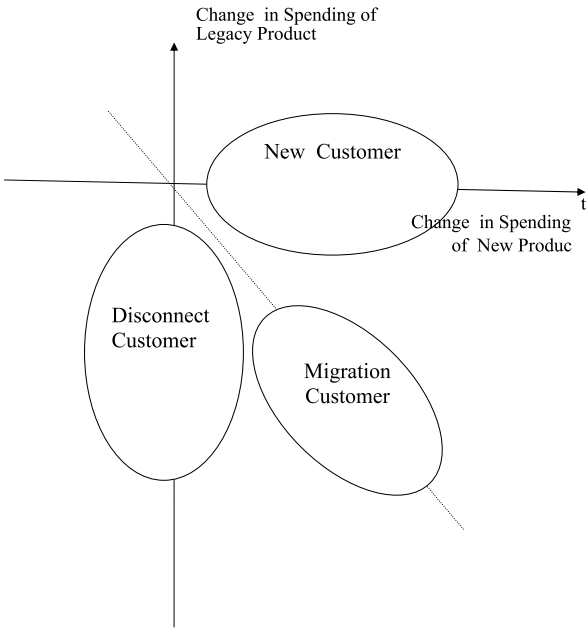
Fig. 1 Customer activities

(b) *Disconnected customers* Figure 1(b) is an example of a disconnected customer whose legacy product revenue was totally lost at certain point within the time frame but has no change of the new product revenue. They may be losses to a competitor and are the type of customers the corporation should put much effort to retain.

(c) *Migrated customers* Figures 1(c), (d) show different examples of the most interesting customers of the paper—migration customers—whose legacy product revenue declined while the new product revenue increased with certain relationships. In general, the lose from the legacy product and the gain from the new product have a strong association in the time points of the changes. Figure 1(c) shows clear migration points and Fig. 1(d) depicts gradual migration. In some cases such as the one shown in Fig. 1(c), the magnitudes of revenue change may be similar but with opposite directions. Both of these two types of migration customers are the focus of this paper since we are interested to know when these changes happened for each customer and how the magnitudes can be related in the sense of the overall revenue change. For the time being, we will not distinguish these two types of customers.

(d) *Others* The rest of customers besides the new, disconnected, and migration corresponds to those with no significant revenue changes in both products. They might be the customers before potential migrations or have already experienced migration from the legacy product to the new product before the time frame considered. There could be many possible revenue patterns for these customers different from above and the patterns are often masked by noises in the billing systems. More importantly, this

Fig. 2 Customer groups



group is the base of the enterprize customers, which is not only of large number but also accounts for large portions of enterprize revenues.

Figure 2 depicts the schematic relationship between revenue changes of the new product and legacy product for the above four types of customers. However, it is generally difficult to distinguish these groups given the large amount of customer spending records due to enormous noises in billing data. Therefore a data mining framework is needed to extract the basic customer patterns to distinguish new, disconnect and migration customers. Since the purpose of the migration analysis is to explore the relationships of the bivariate time series of billing revenue, a statistical model is needed to identify certain matches between the time series patterns. By matching the patterns of the relationships between the legacy and new product revenues, we can group similar customers and score migration customers based on their statistical significance and business importance for migration analysis.

It is important to point out that modeling bivariate time series of the legacy and new product revenue is also complicated by the so-called lagged response of the new product revenue for business customers. That is, migration customers tend to withhold the legacy product for a while to secure the continuity of service after adopting the new product. Thus the change in the legacy product revenue is often later than that of the new product. For example, Fig. 1(c) shows the migration customer whose legacy product revenue drops down at month 20 while the new product revenue jumps at month 18. This is an important feature of business customers, which is different from consumer behaviors.

3.3 A co-integration model for migration analysis

Time series data mining has been used in a wide range of applications such as speech recognition, financial or stock market prediction, medical image activity detection, microarray pattern recognition, etc. Mining time series usually include tasks of (i) index & retrieval (Agrawal et al. 1993; Keogh 2002; Yi et al. 1998), (ii) classification (Geurts 2001;

Pavlovic et al. 1999; Petridis and Kehagias 1997), (iii) clustering (Corpet 1997; Smyth 1997; Xiong and Yeung 2002). Keogh et al. (2002) and Roddick et al. (2002) present an excellent survey of time series data mining. Most of time series data mining methods target on univariate time series. There have been very few approaches that deal with understanding the relation between two or more time series. For multiple time series, Vector Auto Regressive models (VAR) and Vector Auto Regressive Moving Average models (ARMAV) have been widely used in econometric analysis (Abraham 1980). However these methods rely on stationary assumptions while most of business and economic time series are indeed realizations of nonstationary processes. Granger (1981) introduces the concept of co-integration for nonstationary multivariate time series. Engle and Granger (1987) develop a breakthrough on co-integration analysis to capture the notion that nonstationary variables may possess long-run equilibrium relationships so that they tend to move together in the long run. The Appendix provides an overview of the co-integration analysis in econometrics.

Here we assume a time series database consisting of N customers each with T time periods of billing revenue history. Denote $(X_i, Y_i) = \{x_i(t), y_i(t)\}$ the billing revenue of new and legacy products, respectively, for each customer, where $i = 1, \ldots, N$ and $t = 1, \ldots, T$. The value of the time series could be dollars spent, telephone usage, store or web visits, etc. More interestingly, the two time series are additive, which describe the overall spending/usage/visits for each customer. Our objective is to mine the time series database to explain the changes of the overall revenue of each product, $X(t) = \sum_{i=1}^{N} x_i(t)$ and $Y(t) = \sum_{i=1}^{N} y_i(t)$. In particular, we are interested in quantifying the association between the bivariate time series in order to identify important customers that contribute to the revenue changes significantly.

The properties of migration customers discussed above imply a stable relationship among the bivariate time series of the legacy and new products. For simplicity, a linear regression model can be built between X_i and Y_i as follows,

$$Y_i = X_i \beta_i + \varepsilon_i \tag{1}$$

where β_i is the regression coefficient measuring the relationship between Y_i and X_i and ε_i is a zero-mean noise. Although the linear regression model in (1) is attractive to capture the relationship between X_i and Y_i, Granger Representation Theorem (Granger 1981) shows that there may be multiple solutions of β_i if X_i and Y_i are not stationary. Only when ε_i is $I(0)$, which means 0-degree stationary, can β_i be uniquely determined, which carries insights of business relationships. In this case, the variables X_i and Y_i are called co-integrated.

In order to test if ε_i is stationary or $I(0)$, a number of statistical tests have been developed. A brief discussion of these tests can be found in the Appendix. Here we use Engle-Granger test due to its popularity in economic analysis. The co-integration test allows us to assess stable relationships in the migration analysis since it is expected that migration customers tend to maintain a more stable relationship than new and disconnected customers and also the migration customers have different characteristics of co-integration coefficient than the rest of other customers.

3.3.1 Extended co-integration test

In order to accommodate the delay response in the above model, we extend the standard co-integration model to handle lagged responses between the two revenue time series. The extended co-integration model (ECIT) uses a sliding window of τ $(0 \leq \tau \leq P)$ between X_i and Y_i, where τ is an integer that represents potential lagged response of new product. $P = 3$

was specified by marketing experts and used in the following case study. That is, we assume multiple possible linear relationships as follows,

$$Y_{i\tau} = X_{i\tau}\beta_{i\tau} + \varepsilon_{i\tau}, \quad 0 \le \tau \le P \tag{2}$$

where $X_{i\tau} = (x_i(\tau+1), x_i(\tau+2), \ldots, x_i(T))'$ and $Y_{i\tau} = (y_i(1), y_i(2), \ldots, y_i(T-\tau))'$. Using the least squares method, one can estimate the co-integration coefficient $\hat{\beta}_{i\tau}$ and obtain the residual vector $\boldsymbol{e}_{i\tau} = Y_{i\tau} - \hat{Y}_{i\tau}$, where $\hat{Y}_{i\tau} = X_{i\tau}\hat{\beta}_{i\tau}$. The corresponding goodness-of-fit is measured by

$$R^2_{i\tau} = 1 - \frac{\boldsymbol{e}'_{i\tau}\boldsymbol{e}_{i\tau}}{(Y_{i\tau} - \bar{Y}_{i\tau})'(Y_{i\tau} - \bar{Y}_{i\tau})}, \tag{3}$$

where $\bar{Y}_{i\tau} = \sum Y_{i\tau}/(T-\tau)$.

We then apply the ADF test statistic (see Appendix), an autoregressive unit root test, to test the stationarity of residuals $\boldsymbol{e}_{i\tau}$. The best lag τ^* of the delayed relationship is selected with the maximum values of $R^2_{i\tau}$ among those that reject the unit root test. Consequently, the best fit co-integration coefficient is $\beta_i = \beta_{i\tau^*}$. Otherwise, if no τ in the lagged relationships retains co-integration relationships, we set $\beta_i = 0$.

3.3.2 Migration index

It is important to note that the above test considers only statistical significance of the time series patterns while the customer business values are ignored. In practice, maximizing profits is the key driver for service/product providers, and customers couldn't be considered the same even though their patterns may be matched significantly in the statistical test. Business values have to be extracted from the revenue time series in order to prioritize the migration efforts. To enhance the power of different statistical tests and, more importantly, consider the business impacts of different types of customers, we define a migration index for each customer so that customers can be ranked based on their business importance in terms of product/service migration.

First, since companies concern about revenue changes due to product/service migration, migration revenue impact is defined as the minimum of the revenue increase of the new product and the revenue decrease of the legacy product for each customer, i.e.,

$$\alpha_i = \min(\Delta X_i, -\Delta Y_i). \tag{4}$$

The minimum allows us to pick customers having the most significant revenue changes. In the co-integration analysis of the relationships between X_i and Y_i, the coefficient β_i not only measures the likelihood of the migration, but also the strength of the product/service substitution. The larger $|\beta_i|$ is, the stronger the substitution. Therefore a migration index for customer i is defined as follows to quantify the risks of migration,

$$M_i = |\beta_i| * \alpha_i. \tag{5}$$

The migration index M_i combines the likelihood and revenue impact of the migration for customer i.

To screen out the most important migration customers, a threshold T_M is set by business experts and a significant migration customer is identified when

$$\begin{cases} \texttt{Customer } i \texttt{ migrates} & \text{if } M_i \ge T_M \\ \texttt{Customer } i \texttt{ not migrates} & \text{if } M_i < T_M \end{cases}. \tag{6}$$

This migration index allows the corporation to prioritize the marketing efforts to targeted customers for retention and product cross-selling. Next we will use an industrial example to illustrate the application of this data mining framework for migration analysis.

4 An industrial example

As discussed in Sect. 2, the telecommunication market is experiencing a fast-paced development in both technology advancement and fierce competition. Various IP applications have put forward critical challenges to existing communication technologies and brought enormous opportunities in business development. We now apply the above ECIT model to identify migration customers in a telecommunication company to assist marketing decisions when introducing new product. Here the legacy product may refer to packet and related services and the new product may refer to IP or other related products. As a benchmark, we compare the ECIT model with the traditional change point detection method based on generalized likelihood ratio test (GLRT) for identifying migration customers.

4.1 Data description and preprocessing

The dataset that involves in this migration analysis is a sample of 2000 customers with consecutive 36 months of revenue billing history for both legacy and new products. As illustrated in Fig. 3, customer spending time series is always noisy and frequently disrupted by various customer activities such as spending trends, seasonal effects, account adjustments,

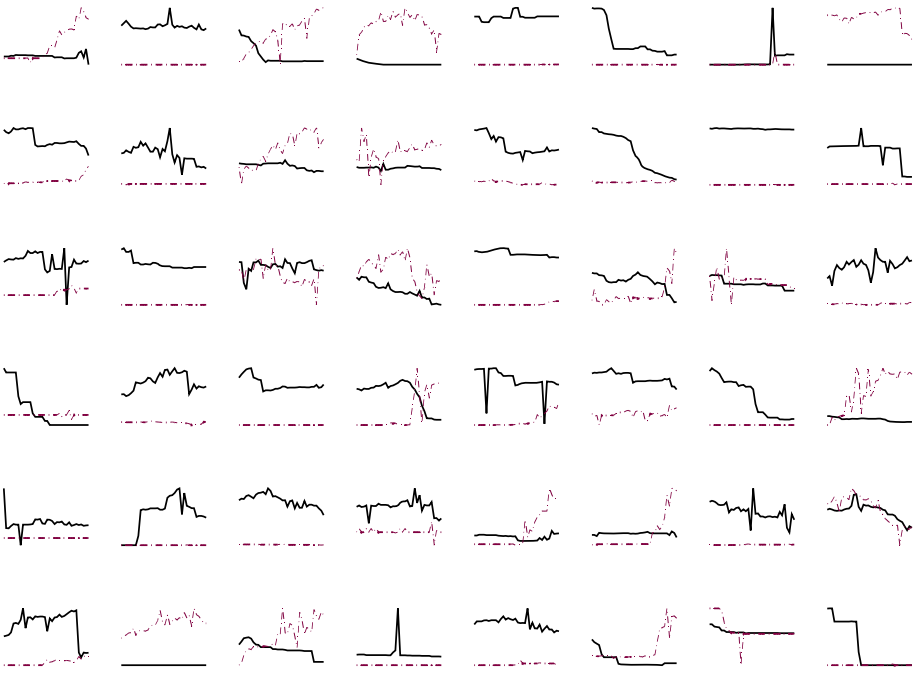


Fig. 3 Original customer spending time series. (*Solid line* represents the legacy product and *dashed line* represents the new product)

contract renews, business development, and billing errors. These time series patterns are so complex and noisy that the standard statistical hypothesis testing methods may fail to capture the true relationship between the two products without cleaning and pre-processing the data. It is therefore important to filter out attributes that are irrelevant to the migration study, e.g., outliers which are out of our interests since they are not a feature for migration analysis but could seriously bias the estimation results. In order to alleviate the effects caused by outliers, we apply 3-month moving median filter for pre-processing and smoothing the data while retaining trend and change point as key migration features in the analysis. Since seasonal effects need to be considered for pre-processing, we also apply centered moving averages to de-seasonalize each time series.

4.2 Change-point detection method for migration identification

Another naive method for migration identification using time series is change point detection. Theoretically, if change points can be identified for each time series independently and their locations are compared against each other, the potential product substitution can be identified. Consider a time series x_t which can be described by a parametric family of probability density function $p(x, \theta)$ and θ_0 and θ_1 are the parameters before and after the change. Assuming the change point is at position m $(1 \le m \le T)$, the parameters θ_0 and θ_1 can be estimated directly from the data by the maximum likelihood principle within their respective regions of data $\{1 \cdots m-1\}$ and $\{m \cdots T\}$, i.e.,

$$\hat{\theta}_0 = \arg\max_{\theta} p(X_1^{m-1}, \theta)$$

and

$$\hat{\theta}_1 = \arg\max_{\theta} p(X_m^{T}, \theta)$$

where $X_1^{m-1} = (x_1, x_2, \ldots, x_{m-1})'$, $X_m^T = (x_m, x_{m+1}, \ldots, x_T)'$, and $p(X_m^n, \theta)$ is the joint likelihood of vector $(x_m, x_{m+1}, \ldots, x_n)'$. Similarly, the null hypothesis parameter θ_0 can be estimated from data in the whole region $\{1, \ldots, T\}$, i.e., $\hat{\theta} = \arg\max_{\theta} p(X_1^T, \theta)$.

The generalized (log-)likelihood ratio is defined as follows (Willsky and Jones 1976)

$$\lambda_m = \log \frac{p(X_1^{m-1}, \hat{\theta}_0) p(X_m^T, \hat{\theta}_1)}{p(X_1^T, \hat{\theta})}. \tag{7}$$

However, because the change point time m is unknown, a search must be conducted to locate the most likely change point. The decision function is the maximum of the above likelihood ratio statistic, which is then compared against a preset threshold at each time step. The change point can be estimated as the time index that maximizes the decision function, i.e.,

$$\hat{m} = \arg \max_{1 \le m \le T} \lambda_m. \tag{8}$$

If H_0 is not rejected, we set $\hat{m} = 1$.

Based on the change point detected from (8), assuming that $x_i(t)$ and $y_i(t)$ are both independently and identically distributed, their change points can be estimated as $\hat{m}_x$ and $\hat{m}_y$, respectively, and the change of $x_i(t)$ and $y_i(t)$ are expressed as

$$\begin{cases} \Delta X_i = \frac{\sum_{t=k+1}^{T} x_i(t)}{T-k} - \frac{\sum_{t=1}^{k} x_i(t)}{k}, & k = \hat{m}_x, \\ \Delta Y_i = \frac{\sum_{t=k+1}^{T} y_i(t)}{T-k} - \frac{\sum_{t=1}^{k} y_i(t)}{k}, & k = \hat{m}_y. \end{cases} \tag{9}$$

As discussed before, the migration customers should have positive ΔX_i and negative ΔY_i. Moreover, the change point for legacy product should be no earlier than that of new product, i.e., $\hat{m}_x > \hat{m}_y$. By comparing the values of $\hat{m}_x$ and $\hat{m}_y$ and the magnitudes of the changes, we can score customer and screen migration customers with a pre-defined threshold. Similar as the ECIT method, we define the migration score for customer i as α_i in (4).

Although the GLRT statistic is easy to calculate and statistically efficient in change point detection, the assumption that there is only one single level shift is too restrictive. In reality, especially for migration customers, it is common to have unstable spending with relatively large volatilities and more complicated patterns than level changes, e.g., ramp changes. Most importantly, the GLRT method doesn't impose any relationship between the two revenue times series whereas the ECIT method identifies stable relationships. In the following, we shall compare the ECIT and GLRT methods for migration identification in the telecommunication example.

4.3 Numerical results

4.3.1 Migration groups

We now apply the GLRT and ECIT methods to the revenue time series to identify migration customers and evaluate their business impacts. As an example, Fig. 4 shows scatter plot of the revenue changes of the new and legacy products for the migration customers identified by the two methods using the same revenue change threshold. The threshold is determined by marketing experts. The cross labeled points are 84 migration customers identified by the ECIT method and the circle labeled points are 297 customers identified by the GLRT method. In fact, the customers identified by the ECIT method is a subset of those identified by the GLRT method.

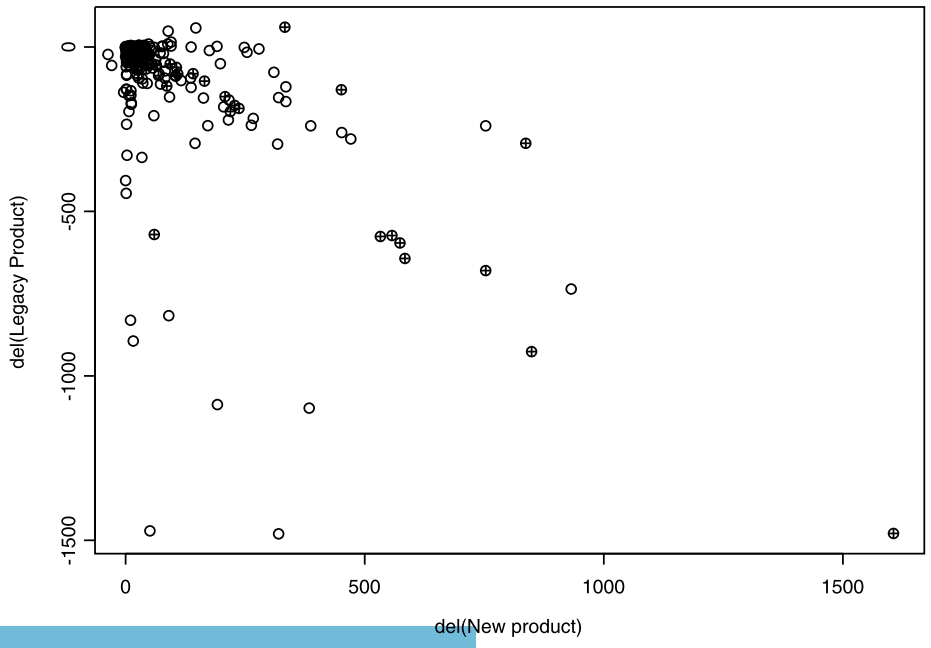


Fig. 4 Migration customers identified by the ECIT and GLRT methods

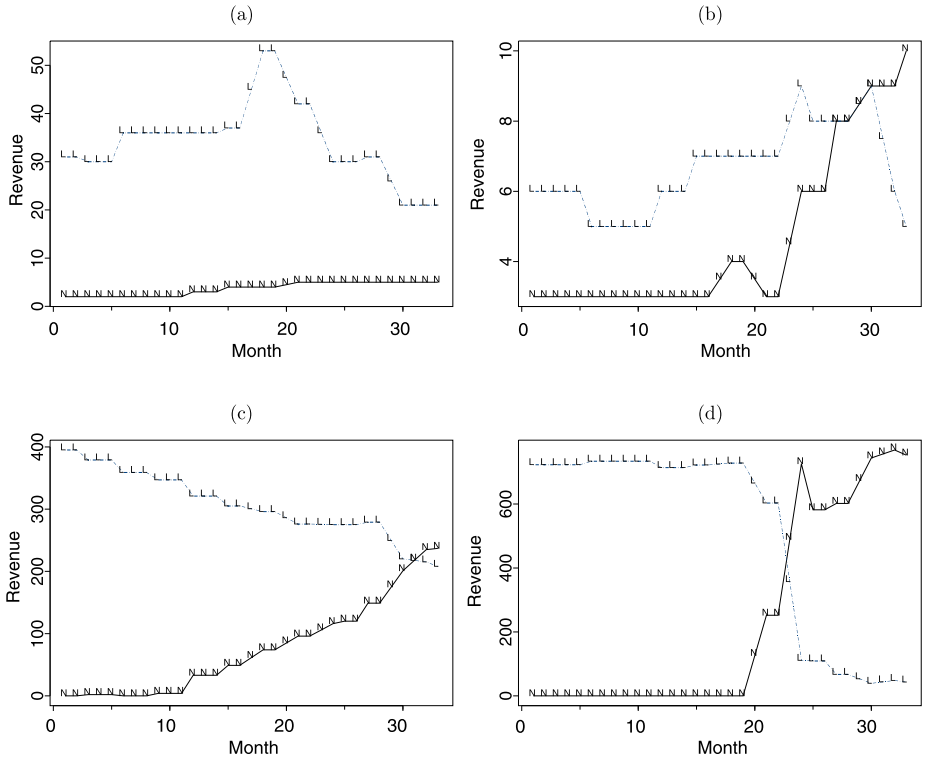


Fig. 5 Customers identified by ECIT and GLRT

However, further investigation shows that the GLRT method produces too many false positives in the identification. For illustration, Fig. 5 presents two examples that are identified by each method. Plots (a) and (b) are identified by the GLRT method and (c), (d) are identified by the ECIT method. Both (a) and (b) represent customers with complicated behaviors for each product. For example, as shown in (a), the revenue of the legacy product declines comparing the starting and end points of the time frame, and the change point of the legacy product is behind the change point of new product. However, it is seen that the revenue of the legacy product actually increases from time 11 to 21, which coincident with that of the new product. A significant drop of the legacy product starts from time 22, but the new product is stable from that time. It is clear the decline of legacy product is not related to the increase of new product in this case. Similarly as shown in (b), the legacy product grows after the new product is introduced and the decline of legacy product seems not related to the new product.

On the other hand, plots (c) and (d) in Fig. 5 present the customers that are interesting to us. The growth of the new product is strongly related to the decline of legacy product and the change points are related to product substitutions. These four examples indicate that the GLRT method may identify more migrating customers with high false positive rate. Although these are only a few example, in the following we shall perform a rigorous analysis of relative operating characteristics (ROC) for the two identification methods.

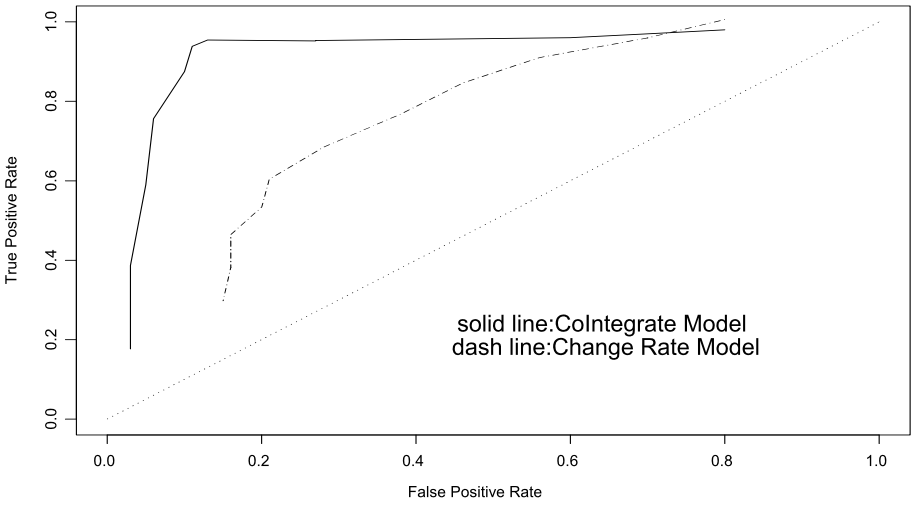


Fig. 6 ROC curve—migration customer counts

Table 1 Revenue-based error type table

	Migration	No-migration
Positive result	*A*. Σ`True positive revenue change`	*B*. Σ`False positive revenue change`
Negative result	*C*. Σ`False negative revenue change`	*D*. Σ`True negative revenue change`

4.4 ROC analysis

To further assess the GLRT and ECIT methods, migration customers have to be labeled for classification analysis. Here we randomly selected 200 customers and sent to marketing experts for labeling and verification of migration customers. In general, the true positive rate (TPR) and false positive rate (FPR) can be assessed by the instance counts and plotted in ROC curves as shown in Fig. 6. If the instance is positive (migration) and it is classified as positive (migration), it is counted as a true positive; if the instance is negative and it is classified as positive, it is counted as a false positive. It is clear from Fig. 6 that, given the same level of false positive rate, the ECIT method often produces a higher true positive rate, i.e., more right classifications. In the other word, the GLRT has a higher false positive rate than the ECIT given the same true positive rate.

As discussed before, it is not only the number of migration customers of great importance to business decisions, but also the revenue that the migration customers represent. To assess the revenue impact of migration customers using the ROC concept, we define the following migration scores in terms of revenue in Table 1 for different types of classifications. False positive and true positive scores weighted by the absolute value of legacy product revenue change are calculated after identifying the migration customers. These scores carry more business meanings than the probability rate presented in Fig. 6. Figure 7 shows the revenue weighted ROC curves for the 200 customers. It is remarkable that the ECIT method is far

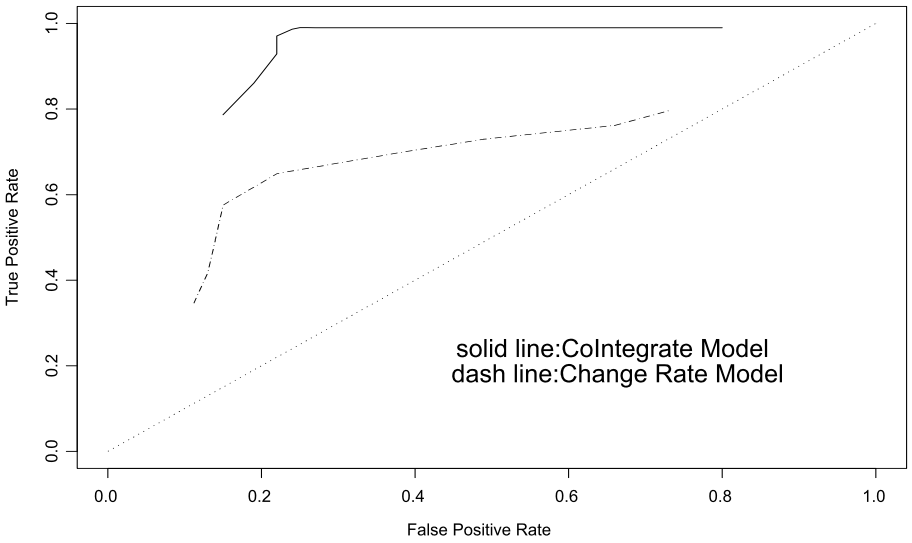


Fig. 7 ROC curve—migration customer revenue

superior than the GLRT method in terms of identifying the right customers with higher revenue impact.

5 Discussion and future research

In this paper, we discuss the product migration problem in competitive market and propose a framework for migration identification and quantification using econometric models and data mining techniques. We elaborate the proper business data that should start with for migration analysis in practice and propose an extended co-integration test to quantify customer migration. More importantly, business knowledge is integrated in the model when defining the migration index and ranking the importance of individual customer. At the end we present an example from the telecommunication industry to illustrate the real world data mining process for migration analysis and compare the co-integration model with the generalized likelihood ratio test. The numerical results have shown that the co-integration model is more effective than the GLRT in both statistical significance and business revenue impact.

Although the proposed co-integration model has shown some encouraging results in migration identification, more complicated co-integration structures may exist in customer billing time series data, e.g., economic structural breaks which appear to be very common in business database. Although we have proposed a lagged response model to capture dynamics of migrated customers, it will be interesting to develop more general co-integration models that allow structural breaks in the future. Moreover, product migration problem is not limited to two products. In reality, migration might involve multiple legacy products and multiple new products. It is of more practical interest to extend our co-integration method to model complex relationships among multiple products. This will be further pursued in another paper.

Appendix

The co-integration concept, introduced by Granger (1981) and formalized in his paper with Engle (1987), has turned out to be extremely important in the analysis of nonstationary economic time series. It becomes common standard to apply co-integration test on economic time series study. The basic point of Granger Theory is that for two or more nonstationary time series, if the linear relationship is stationary, then these time series are called co-integrated, which means residuals from linear regressions are stationary. Many co-integration tests have been developed incorporating with structure breaks, ramp function, etc. Here we briefly introduce the one we use in the paper and the most popular co-integration test, Engle-Granger test (Faloutsos et al. 1994).

The Engle-Granger (EG) test proceeds in two steps. The first step is simply running the static ordinary least square (OLS) estimation.

$$y_t = \alpha + \beta x_t + \mu_t, \tag{A.1}$$

which captures potential long-run relationship between the variables x_t and y_t. In the second step, Dickey-Fuller (DF) test is applied to test if the residual μ_t is stationary.

For the simple AR(1) model

$$\mu_t = \rho \mu_{t-1} + \epsilon_t, \tag{A.2}$$

where ϵ_t $I(0)$, the DF test stems from the null and alternative hypothesis,

$$\begin{cases} H_0 : \rho = 1 & \text{(has unit-root)} \\ H_1 : |\rho| < 1 & \text{(has root outside unit circle).} \end{cases} \tag{A.3}$$

In particular, the regression model for DF Test is:

$$\Delta \mu_t = (\rho - 1)\mu_{t-1} + \epsilon_t = \phi \mu_{t-1} + \epsilon_t. \tag{A.4}$$

The null and alternative hypothesis change to test if ϕ equals zero. A table has been used to replace standard t-distribution for critical values. The time series y_t and x_t are called co-integrated if the t-statistics for testing the hypothesis $\phi = 0$ is smaller than the corresponding critical value in the DF table. On the other hand, if we cannot reject the hypothesis $\phi = 0$, then there will be a unit root in the residuals, and therefore, the series y_t and x_t are not co-integrated.

There are many extensions and generalization of the DF test in (A.4). When deterministic trend and drift are presented and serial correlations are concerned in the regression residuals, Elliott et al. (1996) propose the Augmented Dickey-Fuller (ADF) Test by adding various lagged dependent variables on the base of DF test,

$$\Delta \mu_t = \omega + \xi t + \phi \mu_{t-1} + \delta_1 \Delta \mu_{t-1} + \cdots + \delta_p \Delta \mu_{t-p} + \epsilon_t \tag{A.5}$$

where p is the lag order of the autoregressive process. The correct value for number of lags can be determined by minimizing information criteria such as the Akaike information criteria (AIC) or Schwarz-Bayesian information criteria (BIC).

References

Abraham, B. (1980). Intervention analysis and multiple time series. *Biometrika*, *67*(1), 73–78.

Agrawal, R., Faloutsos, C., & Swami, A. (1993). Efficient aimilarity search in sequence databases. In *Proc. of the 4th international conference on foundations of data organization and algorithms* (pp. 69–84).

Allenet, B., & Barry, H. (2003). Opinion and behaviour of pharmacists towards the substitution of branded drugs by generic drugs: survey of 1,000 French community pharmacists. *Pharmacy World & Science*, *25*(5), 197–202.

Andrew, A. (2007) Blockbuster Company profile; retrieved from http://topics.nytimes.com/top/news/business/companies/blockbuster_inc/index.html, accessed on 13th, October, 2007.

Base, F. (1969). A new product growth model for consumer durables. *Management Science*, *15*, 215–227.

Bloomberg (2007) Merck profit falls as generic copies hurt zocor sales (Update6). http://www.bloomberg.com/apps/news?pid=newsarchive&refer=home&sid=a4lZ.OCpC5II. Jan. 30, 2007.

Constantiou, I. D., & Kautz, K. (2008). Economic factors and diffusion of IP telephony: empirical evidence from an advanced market. *Telecommunication Policy*, *32*(3–4), 197–211.

Corpet, F. (1997). Multiple sequence alignment with hierarchical clustering. *Nucleic Acids Research*, *16*, 10881–10890.

Elliott, G., Rothenberg, T., & Stock, J. (1996). Efficient tests for an autoregressive unit root. *Econometrica*, *64*(4), 813–836.

Engle, R., & Granger, C. (1987). Co-integration and error correction: representation, estimation, and testing. *Econometrica*, *55*(2), 251–276.

Faloutsos, C., Ranganathan, M., & Manolopoulos, T. (1994). Fast subsequence matching in time-series databases. In *Proc. ACM SIGMOD international conference on management of data* (pp. 419–429).

Geurts, P. (2001). Pattern extraction for time-series classification. In *Proc. of PKDD 5th European conference on principles of data mining and knowledge discovery* (pp. 115–127).

Granger, C. W. J. (1981). Some properties of time series data and their use in econometric model specification. *Journal of Econometrics*, *16*, 121–130.

IBM Co (2008) http://www-03.ibm.com/press/us/en/pressrelease/25737.wss.

Johnson, W., & Bhatia, K. (1997). Technological substitution in mobile communication. *Journal of Business and Industrial Marketing*, *12*(6), 383–386.

Keogh, E. (2002). Exact indexing of dynamic time warping. In *Proc. 28th international conference on very large data bases* (pp. 406–417).

Keogh, E., & Kasetty, S. (2002). On the need for time series data mining benchmarks: a survey and empirical demonstration. In *Proc. 8th ACM SIGKDD international conference on knowledge discovery and data mining* (pp. 102–111).

Mahajan, V., Muller, E., & Bass, F. (1990). New product diffusion models in marketing: a review and directions for research. *Journal of Marketing*, *51*, 1–36.

Meuter, M., Bitner, M., Ostrum, A. L., & Brown, S. W. (2005). Choosing among alternative service delivery modes: an investigation of customer trial of self-service technologies. *Journal of Marketing*, *69*(2), 61–83.

Murynets, I., Ramirez-Marquez, J., & Jiang, W. (2009). Migration of customers due to service substitution. *Journal of Targeting, Measurement and Analysis for Marketing*, *17*(4), 297–306.

Pavlovic, V., Frey, B., & Huang, T. (1999). Time-series classification using mixed-state dynamic Bayesian networks. In *IEEE computer society conference on computer vision and pattern recognition (CVPR'99)* (p. 2609).

Petridis, V., & Kehagias, A. (1997). Predictive modular fuzzy systems for time-series classification. *IEEE Transactions on Fuzzy Systems*, *5*(3), 381–397.

Porter, M. (1979) How competitive forces shape strategy. *Harvard Business Review*, March/April, 1979.

Prins, R. (2008) *Modeling consumer adoption and usage of value-added mobile services*. Erasmus Research Institute of Management Ph.D. Series Research in Management.

Reisinger, D. (2008) Opinion: can blockbuster be saved. http://arstechnica.com/news.

Roddick, J., & Spiliopoulou, M. (2002). A survey of temporal knowledge discovery paradigms and methods. *IEEE Transactions on Knowledge and Data Engineering*, *14*(4), 750–767.

Rogers, E. (1976). New product adoption and diffusion. *Journal of Comsumer Research*, *2*, 290–301.

Smyth, P. (1997). Clustering sequences with Hidden Markov models. *Advances in Neural Information Processing Systems*, *9*, 648–655.

TeleGeography (2008) http://www.telegeography.com/wordpress/?m=200805, May 2008.

Willsky, A., & Jones, H. (1976). A generalized likelihood ratio approach to detection and estimation of jumps in linear systems. *IEEE Transactions on Automatic Control*, *21*(1), 108–112.

Xiong, Y., & Yeung, D. (2002). Mixtures of ARIMA models for model-based time series clustering. In *Proc. of IEEE international conference on data mining* (pp. 717–720).

Yankee Group (2007). http://www.yankeegroup.com/link.do?method=getbytaxonomy.

Yi, B., Jagadish, H., & Faloutsos, C. (1998). Efficient retrieval of similar time sequences under time warping. In *Proc. of the 14th international conference on data engineering* (pp. 201–208).