DATA MINING ON TIME SERIES: AN ILLUSTRATION USING FAST-FOOD RESTAURANT FRANCHISE DATA

Lon-Mu Liu, Siddhartha Bhattacharyya, Stanley L. Sclove, Rong Chen
Dept. of Information and Decision Sciences (M/C 294)
The University of Illinois at Chicago
601 S. Morgan Street
Chicago, IL 60607-7124

William J. Lattyak
Scientific Computing Associates Corp.
River Forest, IL 60305

ABSTRACT

With the prevalent use of modern information technology, a large number of time series may be collected during normal business operations. We use a fast-food restaurant franchise as an example to illustrate how data mining can be applied to such time series, and help the franchise reap the benefits of such an effort. Time series data mining at both the store level and corporate level are discussed. Related data warehousing issues are also addressed. Box-Jenkins seasonal ARIMA models are employed to analyze and forecast the time series. Instead of a traditional manual approach of Box-Jenkins modeling, an automatic time series modeling procedure is employed to analyze a large number of highly periodic time series. In addition, an automatic outlier detection and adjustment procedure is used for both model estimation and forecasting. The improvement in forecast performance due to outlier adjustment is demonstrated. Adjustment of forecasts based on stored historical estimates of like-events is also discussed. To illustrate the feasibility and simplicity of the above automatic procedures for time series data mining, the SCA Statistical System is employed to perform the related analysis.

Keywords: Automatic time series modeling, Expert system, Outliers, Knowledge discovery, Forecasting

1. Introduction

The modern economy has become more and more information-based. This has profoundly altered the environment in which businesses and other organizations operate. Hence it has also altered the way in which business operations and business data are analyzed. With the prevalent use of information technology, a large number of data are collected in on-line, real-time environments, which results in massive volumes of data. Such *time-ordered* data typically can be aggregated with an appropriate time interval, yielding a large volume of equally spaced *time series* data. Such data can be explored and analyzed using many useful tools and methodologies developed in modern time series analysis. As retail scanning systems, point-of-sale (POS) systems, and more recently on-line transactions through electronic commerce, become indispensable in business operations, time series data and analysis of such data will also become an integral part of effective business operation.

In this paper, we apply data mining in exploration and knowledge discovery when a large number of time series are available for business applications. As mentioned in Friedman (1997), data mining is at best a vaguely defined field; its definition depends largely on the background and views of the definer. Fayyad (1997) viewed that any algorithm that enumerates patterns from, or fits models to, data is data mining. Fayyad further viewed data mining to be a single step in a larger process of knowledge discovery in databases (KDD). KDD is considered to be a more encompassing process which includes data warehousing, target data selection, data cleaning, preprocessing, transformation and reduction, data mining, model selection, evaluation and interpretation, and finally consolidation and use of the extracted "knowledge". Weiss and Indurkhya (1998) broadly defined data mining as the search for valuable information in large volumes of data. Other researchers more directly tie data mining to pattern or knowledge discovery in large databases, and the predictive ability in using such patterns or knowledge in real-life application (see *e.g.* Glymour, Madigan, Pregibon, and Smyth, 1997, and Hand 1998). Regardless of the viewpoints of individual data miners, it is certain that the scope of data mining and its application will expand more and more.

Time series analysis is often associated with the discovery and use of patterns (such as periodicity, seasonality, or cycle), and prediction of future values (specifically termed *forecasting* in the time series context). Therefore one may wonder what are the differences between traditional time series analysis and data mining on time series. One key difference is the large number of series involved in time series data

mining. As a result, a highly automated modeling approach becomes indispensable in such applications. As shown in Box and Jenkins (1970) and a vast volume of time series literature, traditional time series analysis and modeling tend to be based on non-automatic and trial-and-error approaches. When a large number of time series are involved in data analysis and application, development of time series models using a non-automatic approach becomes impractical. In addition to automatic model building, discovery of knowledge associated with events known or unknown *a priori* can provide valuable information toward the success of a business operation. Therefore outlier detection in time series is an essential component of time series data mining. Some outliers reveal errors; others are not errors but exceptions, representing connections that may be keys to new knowledge and potential opportunities. In addition to the above data mining aspects, we shall discuss temporal aggregation of time series, and its implications in data warehousing of time series. These issues are also important components of time series data mining.

In this paper, we employ a real-life business example to show the need for and benefits of data mining on time series, and discuss some automatic procedures that may be used in such an application. To have a better focus, we shall employ one particular example to illustrate the application of data mining on time series. The concepts and methodologies can be readily applied to other similar business operations. In Section 2, we describe the business operation of this example. After that we present the methodology for data mining and knowledge discovery in time series, with special reference to Box-Jenkins seasonal ARIMA (autoregressive-integrated moving average) models. In this section, automatic procedures for time series modeling, outlier detection, and forecasting with outlier adjustment are presented. In Section 4, additional applications of data mining using the developed methodologies are discussed. Some data warehousing issues for this business operation are addressed. In Section 5, a summary and discussion of this research is presented.

2. An Example of Business Operation and Data Mining Application

In this section, we describe the general operation of a fast-food restaurant franchise and outline how data are collected to support restaurant operations and product planning. In the later sections of this paper, we shall discuss the methodologies and potential application of data mining on time series collected by the individual restaurants and corporate office.

The restaurant franchise to be described is one of the world's largest multi-brand fast-food restaurant chains with more than 30,000 stores worldwide. Taking advantage of recent advancements in information technology, this restaurant franchise has modernized its business operations at the store level using relatively inexpensive PC-based servers, and at the corporate level using highly scalable parallel processing architectures.

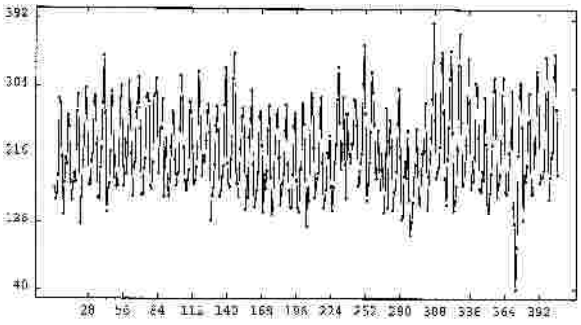
The data collection process involves a POS system. Each time a customer order is placed and the information is keyed into a front register, the transaction is automatically processed through the POS system, time stamped, and then stored in a back-office database. Each restaurant tracks all menu items in addition to the ingredients that go into producing the menu items. This yields several hundred time-ordered series. The centralized corporate office also collects higher level data from each individual restaurant on a regular basis and stores the information in a data warehouse. The sales and transaction data collected by the restaurant chains may be explored and analyzed at both the store level and the corporate level. At the store level, exploring or mining the large amounts of transaction data allows each restaurant to improve its operations management (such as labor scheduling) and product management (such as inventory replenishment and product preparation scheduling), thereby reducing restaurant operating expenses and increasing food quality. At the corporate level, mining pertinent information across the restaurants could greatly facilitate corporate strategic planning. Here, management can assess the impact of promotional activities on sales, evaluate business trends, conduct price sensitivity analysis, and the like. Since the dates and times of the transactions are recorded along with the product information, the data can be easily aggregated into various forms of equally spaced time series. The granularity of the aggregation (*e.g.*, hourly, daily, weekly, *etc.*) is application-specific. For example, if a restaurant needs to know the amount of inventory required on a day to day basis, the data may be aggregated into daily time intervals.

To simplify our discussion, in the next section we shall limit our example to daily time series related to restaurant operations. We shall discuss day-of-week patterns that may exist within the representative data. In addition, we shall discuss the impact of external events (*e.g.*, holidays, local sports events, and outliers) on the modeling and data analysis process. Finally, we shall discuss the necessity of data cleaning methods to reduce distortion in a time series and make a time series less prone to modeling error when applying automated modeling methods.

The graph below depicts the amount of a perishable ingredient used by an individual restaurant for various menu items between April 7, 1997 and May 18, 1998 (a total of 407 observations). The data are aggregated into daily intervals and presented in time order. This daily

time series is in its original form, and no adjustments of any special event effects have been applied to the data. Our primary interest is to forecast the daily demand of this ingredient as accurately as possible in order to facilitate better inventory management. To forecast the demand, a Box-Jenkins univariate time series model will be employed. We are interested in examining this series to exploit homogeneous patterns for forecasting. We are also interested in identifying data that deviates from the expected patterns of the series, which may be caused by events known or unknown *a priori*. By accounting for this information in the form of a model, and by identifying appropriate models in an automated fashion, we are performing analysis on time series data that will assist management is a variety of activities (the most obvious being inventory management). For the study to be presented in the next section, we shall only use the first 365 days (one year) of data for analysis; the remaining 42 days of data are used for comparison of forecast performance.

Figure 1. Daily demand of a perishable ingredient for a fast-food restaurant (4/7/97 ~ 5/18/98)



3. Methodology for Data Mining and Knowledge Discovery in Time Series

In this section we shall discuss time series data mining at the store level. Some of the methodology discussed can also be used in data mining at the corporate level. The time series plot shown in Figure 1 reveals that the series is highly periodic (or seasonal); however it is difficult to see the pattern within each period. In Figure 2, we display the median daily demand from Monday to Sunday using the first 365 days of data. In this plot, we observe that the demand increases from Monday through Saturday (with Saturday similar to Friday), and then drops on Sunday. Instead of using a bar plot, the display in Figure 2 can be replaced by a box-and-whisker plot as shown in Figure 3. In addition to the median of daily demand, a box-and-whisker plot provides information on dispersion (through the use of quartiles and whiskers) as well as the outliers (indicated by "*" in the plot) for each day of the week; thus the characteristics of the weekly pattern are better revealed. In Figure 3, we observe the same median daily demand pattern from Monday to Sunday as in Figure 2. However, while the median daily demands for Friday and Saturday are similar, we observe that Friday has a more disperse distribution for the demand below the median and Saturday has a more disperse distribution for the demand above the median. Even though box-and-whisker plots are more informative for statistically trained personnel, they may be too overwhelming for a typical restaurant manager whose primary concern is the overall operation of the restaurant and who has to handle a large number of time series on daily basis.

Figure 2. Median daily demand (Monday through Sunday) of a perishable ingredient

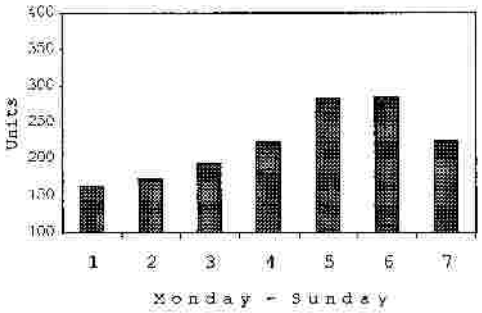
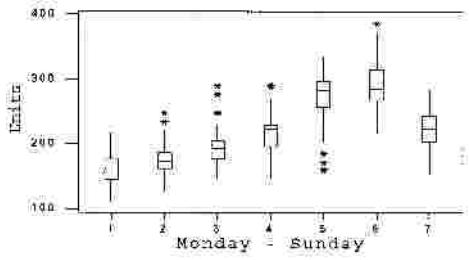


Figure 3. Box-and-whisker plot (Monday through Sunday) for the daily demand of a perishable ingredient.



Although histograms and box-and-whisker plots provide useful insights on the weekly demand patterns of this particular ingredient, it is much more desirable to automatically forecast the daily demand of this ingredient, particularly since in this case a large number of ingredients need to be tracked. To attain accurate forecasts, we first need to develop an adaptive model that has the flexibility to accommodate the versatility of the data. As demonstrated in a number of places in the literature, Box-Jenkins seasonal ARIMA models are very useful in capturing the behavior of seasonal time series and generating accurate forecasts for such series. If nothing else, the ARIMA-based forecasts should be used as a benchmark if other competitive models or methods are to be employed for forecasting comparison. In addition to regular periodic patterns, we also need to develop strategy to capture and incorporate *special events* into the forecasts. Such special events may include

known holidays or festivals, sports activities, and other scheduled local events, *etc.* In this section, we first focus on the development of Box-Jenkins ARIMA models using an automatic approach. Other issues related to forecast accuracy shall be addressed later.

3.1 Box-Jenkins Seasonal ARIMA Models

Using the backshift operator "B" (where $B\,Y_t = Y_{t-1}$), a general multiplicative seasonal ARIMA model can be expressed as

$$\phi(B)\Phi(B^s)(1-B)^d(1-B^s)^D Y_t = C_0 + \theta(B)\Theta(B^s)a_t\,, \quad t = 1,2,\ldots,n \qquad (1)$$

or alternatively

$$(1-B)^d(1-B^s)^D Y_t = C + \frac{q(B)\Theta(B^s)}{f(B)\Phi(B^s)}a_t \qquad (2)$$

where $\{Y_t\}$ is a time series with n observations, $\{a_t\}$ is a sequence of random errors that are independently and identically distributed with a normal distribution $N(0, \sigma_a^2)$, "d" and "D" are the orders of non-seasonal and seasonal differencings for the time series, "s" is the seasonality or periodicity of the series, and $\phi(B)$, $\Phi(B^s)$, $\theta(B)$ and $\Theta(B^s)$ operators are polynomials in B with the following general forms

$$\phi(B) = 1 - \phi_1 B - \phi_2 B^2 - \ldots - \phi_p B^p\,,$$

$$\Phi(B^s) = 1 - \Phi_1 B^s - \Phi_2 B^{2s} - \ldots - \Phi_P B^{Ps}\,,$$

$$\theta(B) = 1 - \theta_1 B - \theta_2 B^2 - \ldots - \theta_q B^q\,,$$

$$\Theta(B^s) = 1 - \Theta_1 B^s - \Theta_2 B^{2s} - \ldots - \Theta_Q B^{Qs}\,.$$

In the above models, the polynomials $\phi(B)$ and $\theta(B)$ capture the non-seasonal behavior of the series, and $\Phi(B^s)$ and $\Theta(B^s)$ capture the seasonal behavior of the series. The differencing orders "d" and "D" typically have a value of 0 or 1, and seldom greater than that. Depending upon the values of the differencing orders, the constant term "C" in (2) may represent the mean, the first-order trend, or the higher-order trend of the time series, while the term "C_0" in (1) does not have specific meaning. As discussed in Liu (1993, 1999), the latter expression (2) therefore is more comprehensive and easier to interpret than the traditional expression in (1) when "C" and "C_0" are non-zero. We shall use the second expression of an ARIMA model in this paper.

To develop an appropriate model for forecasting, Box and Jenkins (1970) employed an iterative procedure involving (a) identification of a tentative model, (b) estimation of model parameters, and (c) checking the adequacy of the tentative model. This iterative procedure requires visual examination of intermediate statistical results, and the model is eventually developed based on the expert judgement of the analyst. Liu (1993, 1999) developed an approach for automatic modeling of non-seasonal and seasonal time series using ARIMA models. Unlike most of automatic ARIMA modeling procedures that tend to malfunction in the identification of *seasonal* time series models, this approach performs particularly well in such a circumstance. In addition, Chen and Liu (1993a) developed a procedure for automatic detection of outliers in a time series, and joint estimation of outlier effects and model parameters. By combining these two procedures, the entire Box-Jenkins iterative modeling approach can be automated and greatly simplified, and the need for visual examination of statistics in intermediate analysis can be greatly reduced or eliminated. Since both automatic procedures are available in the SCA Statistical System (Liu and Hudak, 1992, and Liu, 1999), we shall use the commands in this software system to illustrate the simplicity of automatic time series modeling, and their usefulness in time series data mining. Below we revisit the Box-Jenkins modeling approach and provide a brief review of the methods often used in the traditional (manual, non-automatic) approach. In the process, we then illustrate how these tasks can be easily accomplished by the automatic procedures discussed above.

3.2 Model Identification

In the Box-Jenkins iterative modeling approach, model identification proves to be the most complicated and difficult task, particularly if the time series is seasonal or periodic. A number of methods have been developed for manual identification of non-seasonal time series, including using autocorrelation function (ACF), partial autocorrelation function (PACF), extended autocorrelation function (EACF, Tsay and Tiao, 1984), and smallest canonical correlation table (SCAN, Tsay and Tiao, 1985). These methods are useful for non-seasonal time series, but found to be ineffective for seasonal time series. Liu (1989) employed a filtering method for the identification seasonal time series. This method was incorporated in the SCA Statistical System (Liu and Hudak, 1992, and Liu, 1999) and found to be very effective for automatic identification of ARIMA models for both seasonal and non-seasonal time series. In the implementation of this automatic model identification procedure, heuristic rules and expert knowledge are employed in conjunction with the main algorithm in order to delineate certain ambiguities in model identification. Therefore it is more appropriate to regard the actual implementation of the automatic modeling procedure in the SCA System as an *expert system*, rather than just a straightforward *automatic procedure*. Reilly (1980) and Reynolds, Mellichamp, and Smith (1995) also developed automatic methods for identification of ARIMA models for time series. However the method developed by Reynolds, Mellichamp and Smith (1995), which

employed a neural network approach, is restricted to non-seasonal time series, and the method developed by Reilly (1980) works well for non-seasonal time series, but much less satisfactorily for seasonal time series.

In time series data mining, we often need to handle a large number of time series automatically and effectively. With this in mind, it is an absolute necessity to use an efficient automatic method for model identification. By using the automatic model identification command IARIMA of the SCA Statistical System (in this case, the exact SCA command is "IARIMA Y. SEASON 7." with Y the name of the series, and 7 the potential periodicity of the series), the following model and parameter estimates are obtained.

$$(1 - B^7)Y_t = \frac{1 - \Theta_1 B^7}{1 - \phi_1 B} a_t \qquad (3)$$

$$\hat{\phi}_1 = 0.4386 \quad (t = 9.17) \quad \hat{\Theta}_1 = 0.8561 \quad (t = 30.19) \quad \hat{\sigma}_a = 28.7$$

In addition to identifying the model for a time series, the IARIMA command provides estimates for the model parameters and checks the sample ACF of the residuals. Thus in essence the IARIMA command completes the tasks of model identification, parameter estimation, and certain aspects of diagnostic checking in Box-Jenkins iterative modeling procedure. In this case, no significant sample autocorrelations of the residual series are found, so the identified model is adequate. The above model is consistent with that identified manually using the sample ACF and PACF of the series.

3.3 Model Estimation, Outlier Detection, and Knowledge Discovery

The parameter estimates provided by the IARIMA command are based on a conditional maximum likelihood method discussed Box and Jenkins (1970). For time series with strong seasonality and shorter length, it is advisable to estimate model parameters using the exact maximum likelihood method to gain efficiency in parameter estimates (see *e.g* Hillmer and Tiao 1979). Relatively speaking, the exact maximum likelihood algorithm requires much more computing time than the conditional algorithm. However the computing power of modern hardware has made this distinction an insignificant issue. The parameter estimates for the above model based on an exact maximum likelihood method are:

$$\hat{f}_1 = 0.4437 \quad (t = 9.33) \qquad \hat{\Theta}_1 = 0.9294 \quad (t = 40.85) \qquad \hat{s}_a = 27.45$$

The above results show that the estimate of the seasonal moving average parameter (Θ_1) is larger when an exact maximum likelihood method is used, and the residual standard error is somewhat smaller under such a situation. The estimates of the regular autoregressive parameter (ϕ_1) are similar for the exact and conditional methods. The above differences in the results of estimates are consistent with theoretical studies.

Depending upon the behavior of each time series, outlying data in a series potentially could have rather significant impact on the estimates of the model parameters. In addition, outliers in a time series may indicate significant events or exceptions, and provide useful knowledge for the management and operation of a restaurant. In most of the studies (see *e.g.* Fox, 1972, Chang, Tiao, and Chen, 1988, and Tsay, 1988), model estimation and outlier detection (and the consequent outlier adjustment) are conducted in separate steps. Chen and Liu (1993a) developed a joint estimation method that allows for outlier detection and simultaneous estimation of both model parameters and outlier effects in a combined procedure. This capability is available in the SCA System through its OESTIM command. Using the OESTIM command with the critical value 3.5 (*i.e.* the t-value for an outlier estimate) as the criterion for the determination of outliers (Chen and Liu, 1993a, and Liu and Hudak, 1992), the estimates of model parameters and outlier effects for this time series are listed below.

$$\hat{f}_1 = 0.4571 \quad (t = 9.52) \qquad \hat{\Theta}_1 = 0.9513 \quad (t = 45.40) \qquad \hat{s}_a = 22.53$$

Summary of detected outliers, and their types and estimates

TIME	ESTIMATE	t-VALUE	TYPE
89 (07/04/97)	-149.612	-6.64	IO
234 (11/26/97)	76.812	3.79	AO
236 (11/28/97)	-112.896	-5.56	AO
264 (12/26/97)	-98.645	-4.66	TC
269 (12/31/97)	109.350	5.38	AO
285 (01/16/98)	-95.697	-4.51	TC
307 (02/07/98)	88.949	3.95	IO
349 (03/21/98)	-76.771	-3.60	TC

From the above results, we find the estimated residual standard error is much smaller than the previous values, and the parameter estimates for the model are somewhat larger than the previous estimates. Within the training data (*i.e*. the first 365 observations) used for model estimation, 8 outliers are detected. The time periods, the estimates, the t-values and the types of these outliers are listed in the above table. Typically a larger critical value for outlier detection is employed during joint estimation of model parameters and outlier effects so that the parameter estimates will not be biased (Chen and Liu, 1993a). With this in mind, smaller effects due to special events or holidays may not be revealed. To uncover such effects, we may perform additional outlier detection with a smaller critical value but with fixed model parameter estimates obtained in the preceding step.

In time series outlier detection and estimation, four basic types of outliers are typically considered (Chang, Tiao, Chen, 1988, and Tsay, 1988). These are additive outlier (AO), innovational outlier (IO), temporary change (TC), and level shift (LS). Other types of outliers usually can be expressed in combinations of these four basic types. An additive outlier indicates an event that affects a series for one time period only. In regression analysis, typically this is the only type of outlier considered. Unlike an additive outlier, an innovational outlier indicates an event with its effect propagating according to the ARIMA model of the process. In this manner, an IO affects all values observed after its occurrence. A level shift is an event that affects a series at a given time, and its effect becomes permanent afterward. Finally, a temporary change outlier indicates an event having an initial impact which then decays exponentially. More details regarding the mathematical formulation of models for these outliers and their meanings can be found in Liu and Hudak (1992), and Chen and Liu (1993a). In the above outlier summary table, we found that the first outlier (at t=89) could be attributable to the July 4 (Friday) long weekend. The second (t=234) and the third (t=236) outliers were related to Thanksgiving Day (Thursday, November 27). The fourth (t=264) and the fifth (t=269) outliers were associated with Christmas and New Year Eve. The last three outliers (t=285, 307, and 349) could not be attributed to known events in calendar, and could be related to local events or weather conditions. It is useful to note that the above outliers and their types cannot be identified simply by visualization of the time series plot shown in Figure 1.

The benefits of time series outlier detection and estimation do not only provide better model estimates theoretically. More importantly, as shown in this example, outlier detection often leads to discovery of events that may provide useful information or knowledge. Additional interesting examples can be found in various articles (*e.g.* Chang, Tiao, and Chen, 1988, Liu and Chen, 1991, and Chen and Liu, 1993b). From the management point of view, outlier detection is most useful if it is provided in an ongoing basis so a manager can take advantage of the discovered knowledge in the normal course of business operation. This is particular relevant to forecasting, as outliers occurring at the end or near the end of a time series have the most significant impact on forecasts.

3.4 Diagnostic Checking

In diagnostic checking of an estimated model, our primary interests include (a) examining potential lack of fits; and (b) checking if the assumptions of the model are satisfied. If the model is adequate and no lack of fit is present in the estimated model, the sample ACF of the residual series should follow the pattern of a white noise process (*i.e.*, no autocorrelations of the residual series should be significant). For the assumptions of an ARIMA time series model, typically it is assumed that a_t follows a white noise process, which means (i) a_t's are independent; (ii) $E(a_t) = 0$ for all t; (iii) a_t's follow a normal distribution; and (iv) $E(a_t^2) = \sigma_a^2$ for all t. With this in mind, checking lack of fits in (a) also serves the purpose of checking the first assumption in (b) which is to verify the independence of a_t's. For the second assumption ($E(a_t) = 0$ for all t), a_t will be flagged as an outlier in the OESTIM step or when we examine the residual series. The third assumption (*i.e.*, a_t's are normally distributed) typically is not a major problem, and non-normality very often is associated with outliers in the time series. The fourth assumption (*i.e.*, variance constant over time) usually is not a major concern and can be rectified by an appropriate transformation (*e.g.* logarithm) of the time series.

In reviewing the elements discussed above, we find that if we perform outlier detection and adjustment during model estimation and examine the sample ACF of the residual series afterward, the task of diagnostic checking on the estimated model is completed.

The approach used by the IARIMA command in the SCA System favors parsimonious models, therefore it avoids the potential issues of data dredging. When no adequate parsimonious model can be found, it marks the model found as unsatisfactory and at the same time displays the sample ACF of the residual series. Such a situation does not occur often. When it happens, we need to further examine the characteristics of the time series. By using the IARIMA command in conjunction with OESTIM, the task of automatic time series modeling is greatly simplified and the quality of the resulting model is greatly enhanced.

3.5 Forecasting and Evaluation of Forecast Performance

Once a satisfactory model is obtained, generation of forecasts is an easy process if no outlying data occur at or near the forecasting origin. However when outliers occur at or near the forecasting origin, the task of generating accurate forecasts becomes a lot more complicated (Chen and Liu, 1993b). To evaluate forecast performance, we may employ the root mean squared error (RMSE) for the post-sample period. Note that for evaluation purposes, the post-sample period is used to provide fair cross-validation and avoid the potential misleading impression that the fit is better than it really is due to over-fitting the training series. Typically the RMSE is defined as

$$\text{RMSE} = \sqrt{\frac{1}{m}\sum_{t=1}^{m}(Y_t - \hat{Y}_t)^2}$$

where $\hat{Y}_t$ is the one-step-ahead forecast of Y_t based on an estimated model and *m* is the number of forecasts used in the comparison. Assuming that the estimated model is representative of the forecasting period, the post-sample RMSE should be consonant with the residual standard error (σ_a) of the estimated model. While the gross RMSE defined above is appropriate if no outliers exist, this value may be greatly inflated if any outliers exist during the post-sample period. As a result, comparisons of forecast performance based on the gross RMSEs are often misleading and inclusive (Liu and Lin, 1991). We shall denote the gross RMSE defined above as $RMSE_g$. To obtain better insights into the effects of outlier adjustment on forecasting, Chen and Liu (1993b) further considered three variations of RMSE. To highlight the impact of outliers on the comparison of post-sample forecast performance (but still retain the focus of this study), we shall also compute the $RMSE_r$ as discussed in Chen and Liu (1993b). $RMSE_r$ is the post-sample RMSE computed using the time periods excluding outliers and those immediately following the outliers. This revised RMSE is considered since the forecast cannot be improved at the point where an outlier occurs no matter whether outlier adjustment is employed or not, and the forecast immediately following an outlier is subject to the greatest impact depending whether the preceding outlier type is appropriately identified or not. Unfortunately, the type of outlier cannot be determined by data alone if the outlier occurs at the forecasting origin. Under the definition for $RMSE_r$, we expect that $RMSE_r$ will be a better criterion to judge the forecast performance of a model or method when outliers occur during the post-sample period. Using the parameter estimates obtained under the OESTIM command, the following 8 outliers are detected during the post-sample period. Here we use a smaller critical value 2.5 for outlier detection as suggested in Chen and Liu (1993b), and therefore more outliers are detected.

Summary of detected outliers, and their types and estimates in post-sample

TIME	ESTIMATE	t-VALUE	TYPE
369 (04/10/98)	-54.501	-3.52	AO
371 (04/12/98)	-82.301	-5.00	TC
373 (04/14/98)	-106.246	-6.13	IO
375 (04/16/98)	46.832	2.88	TC
379 (04/20/98)	-50.973	-3.29	AO
388 (04/29/98)	45.843	2.95	AO
398 (05/09/98)	36.951	3.86	LS
400 (05/11/98)	-56.229	-3.15	IO

In the above outlier summary table, the first three outliers (t=369, 371, and 373) were related to Good Friday (April 10) and Easter Sunday (April 12). The next three outliers (t=375, 379, and 388) were relatively smaller, and no known events in the calendar could be attributed to. They could be caused by local events or weather conditions. The last two outliers (t=398 and 400) occurred on the days before and after Mother's Day (May 10), and could be related to this event. As discussed in Chen and Liu (1993b), outliers near the end of a time series could be mis-classified due to lack of data (particularly for LS type of classification), therefore the outlier types at t=398 and t=400 might be changed if more data were available. Based on the above results, we find the largest outlier (which is an IO) occurs at t=373, and the second largest outliers (which is a TC) occurs at t=371. Since we cannot determine the type of an outlier at the forecasting origins without specific knowledge for the outlier, for comparison purposes we uniformly assume that these outliers are all of the same type (IO, AO, TC, or LS). The results of the forecast performance without outlier adjustment (using the regular FORECAST command) and with outlier adjustment (using the OFORECAST command) are listed below.

Summary of forecast performance with and without outlier adjustment

Forecasting Methods	Post-sample RMSE	
	$RMSE_g$	$RMSE_r$
FORECAST (no outlier adj.)	33.527	19.723
OFORECAST/IO	33.993	17.609
OFORECAST/AO	36.807	17.523
OFORECAST/TC	33.958	17.643
OFORECAST/LS	36.346	18.627

From the above results, we find that $RMSE_g$ are greatly inflated in comparison with the residual standard error of the estimated model or $RMSE_r$. The $RMSE_g$'s under OFORECAST with the assumptions that all outliers occurred at the forecasting origins being all IO or TC are smaller than those of AO and LS since the largest two outliers are IO and TC (and in this case, IO and TC have similar behavior for the model under study). The $RMSE_r$ are quite similar under all outlier assumptions except when no outlier adjustment is employed in forecasting at all (*i.e.* the first row), or if the outliers at the forecasting origins are all assumed to be level shift (*i.e.* the last row). The LS outliers have a strong impact on the behavior and forecasts of a time series. This type of outlier should be avoided unless there is a strong reason to consider it. Based on the $RMSE_r$'s in the above table, we find that outlier adjustment does improve the accuracy of forecasts.

3.6 Data Cleaning and Handling of Missing Data

For anomalous data with unknown causes, the incorporation of automatic outlier detection and adjustment procedure can ultimately produce more appropriate models, better parameter estimates, and more accurate forecasts. It is also important to note that anomalous data may have known causes and may be repeated. For example, in the restaurant industry, holidays such as Independence Day and special events such as local festivals tend to have significant impact on sales. The effects associated with such known causes can

be estimated and stored in a database if adequate historical data are available (Box and Tiao, 1975). Since holidays and special events are typically known by management and can be anticipated, the associated effects can be used to adjust the model-based forecasts and thus greatly increase the forecast accuracy. Such improvement of forecast accuracy cannot be accomplished by using an outlier adjustment procedure. In addition to forecast adjustment, the stored event effects can be used to clean historical data if it is desired. We may also study the effects of a specific event over time to understand the impact of the event on the business operation.

Similar to other statistical analyses, missing data must also be addressed in the time series context. For example, a restaurant may close due to extreme weather or major power outage. A special consideration in handling missing data in a time series application is that the missing data cannot simply be omitted from the data series. When missing data occur, these observations must be replaced by appropriately estimated values so that the alignment of data between time periods will not be offset inappropriately. As discussed in Chen and Liu (1993a) and Liu and Hudak (1992), missing data in a time series may be temporary replaced by any rough initial value and further refined by treating it as a potential additive outlier. The OESTIM and OFORECAST commands in the SCA System use such an approach and can directly handle estimation and forecasting of a time series with missing data.

3.7 Data Warehousing at the Store Level

Data warehousing is relatively straightforward at the store level. At this level, the data collected through POS system are aggregated into fractional hour intervals, which in turn can be aggregated into hourly and daily intervals. In this study, we focus our research on time series data mining based on daily data. In some other applications, quarter hour or hourly data may be needed.

In addition to data collected through the POS system, it is useful to record and remark on external events, such as special promotions, local events, and holidays in the database. Such information will allow us to estimate the effect due to each kind of special event, which in turn can be used to improve the accuracy of forecast as discussed above. Once the effects of the external events are estimated, they should be stored in the database jointly with event remarks so the information can be employed easily in the future. It may also be useful to store other external information that may affect the sales and operation of a restaurant, such as daily temperature, rainfall, and snowfall, *etc.* Such information will allow us to conduct further study and refine forecasting models or procedures if needed.

4. Data Mining at the Corporate Level and Its Applications

The issues of data mining and data warehousing at the corporate level for this business operation are much more complex than at the store level, yet the potential benefits can also be much more substantial. Even though modern information technology allows us to store huge amounts of data at a relatively inexpensive cost, the sheer number of stores and the number of time series in each store can make data warehousing a formidable task. At the corporate level it may not be possible to store all data that are potentially of interest. However, any important data (*a posteriori*) that are not warehoused can become costly to reconstruct or obtain at later date. In some situations, no remedial solutions may be available, causing irrevocable impairment to the competitiveness of the business operation. With this in mind, it is important to envision the potential applications of the data to be used at the corporate level, and design a flexible and evolving strategy to warehouse the data. The latter point is of particular importance. Since it is unlikely that we can foresee the needs of all future applications, a flexible and efficient strategy to allow for inclusion of new data series in a database or data warehouse is the best antidote to this potential difficulty.

As mentioned in the previous sections, appropriate choice of granularity in temporal aggregation is essential in successful time series data mining. The methodology developed in Section 3 and its extensions can be employed in most of time series data mining at the corporate level. In this section, we shall discuss a few potential applications of data mining at the corporate level, and use these examples to illustrate the importance of appropriate temporal aggregation. Some issues raised in this section can be important considerations in the design of the database and data warehouse.

4.1 Rapid Evaluation of Promotional Effects

It is very common for a corporate office to sponsor various promotional campaigns at both the national level and the regional level. By successfully increasing awareness of a company and its products through promotional activity (*e.g.*, television, radio, print, and coupon drop, *etc.*), fast-food franchises can potentially reap increased market share in addition to enjoying spurts of increased sales.

Before a major promotional campaign is launched, it is prudent to conduct a "pilot study" on the campaign and other alternatives in a smaller scale in some well-defined regions. We can then evaluate the relative effectiveness of these campaigns using the data collected at the store level within each region. By designing the pilot study appropriately, it is possible to evaluate the short-term

promotional effects due to different campaigns *rapidly* and *accurately* by pooling the data across the stores. In such a study, daily data across the stores may be employed. The intervention models discussed in Box and Tiao (1975) may be used to measure the impact of a specific campaign even though the daily data have a strong 7-day periodicity. To avoid the potential complexity caused by the periodicity in daily data, weekly data may be used. However, longer data span may be needed if weekly data are used. Also it may be difficult to measure the initial effects of each promotional campaign in such a case.

When applying intervention analysis (Box and Tiao, 1975), it is important to note that outlying data must be handled appropriately. Otherwise, insignificant results may be obtained even when the true impact of an intervention is significant (Chen and Liu, 1991). This is due to the fact that outliers in general inflate the variance of a time series process. In some situations, outliers can caused biased or inaccurate results since the intervention effects could be overwhelmed by major outlying data which are influenced by some random special events (*e.g.*, a school bus of children happens to stop at a restaurant to eat after a field trip).

4.2 Seasonality Analysis of Product Sales

In the fast-food restaurant business, it is easy to understand that the sales of certain products are highly seasonal. Such seasonality could be caused by annual weather patterns, major holidays or festivals, or regular occurrences of sport activities, *etc.* Understanding the seasonal patterns for the sales of the products across restaurants in a region allows a company to develop more beneficial strategic plans such as changes of menu, general marketing efforts, and special product promotions. This is of particular importance for a publicly traded corporation as Wall Street does not always interpret the normal seasonal patterns of corporate earnings rationally. A better understanding of the seasonality for product sales can be very useful to help a company achieve its goal for sales and revenue, or at least communicate with the financial community more effectively.

An appropriate time interval for studying seasonal sales patterns of fast-food products can be based on monthly aggregated data. However, since day-of-the-week effects are very prominent for daily time series, a time series generated using the aggregate of regular month can create misleading information for the seasonal patterns since the composition of Monday through Sunday from January to December are not the same from year to year. Furthermore such an aggregation procedure can greatly complicate the model identification and estimation of the time series (Liu, 1980, 1986). To avoid such a problem, we may consider using the time series aggregated based on the so called “retail month”. A retail month consists of four complete weeks in each month; therefore there are 13 retail months in each year. The automatic procedures described in Section 3 can be used to model seasonal monthly time series quite effectively, particularly for the series based on retail month. For a time series based on regular months, the composition of the day-of-the-week in each month must be incorporated into the model (Liu, 1986). Otherwise it often requires a rather complicated and implausible model in order to have a clean ACF for the residual series (Thompson and Tiao, 1971, and Liu, 1986). Furthermore the forecasting accuracy can be severely compromised if day-of-the-week information is not included in the model for such time series.

Instead of using monthly data, we may use quarterly data to study the seasonality of product sales. In such a situation, the irregularity caused by the day-of-the-week effects is minimal and can be ignored. However, a more aggregated time series typically contains less information, and therefore also produces less accurate forecasts. No matter whether monthly or quarterly time series are used for seasonality analysis or forecasting, the more data we have the better. In some corporations, older data are often discarded due to lack of storage space, making it very difficult (if not impossible) to analyze monthly or quarterly time series adequately.

4.3 Performance Analysis of Individual Store or Product

At the corporate level, it can be quite useful to study both the best performing (say top 1%) and the worst performing (say bottom 1%) stores. By exploring the characteristics of these stores, useful information may be obtained, which in turn can be used to improve the performance of the stores in the entire corporation. This can be viewed as a form of "management by exception”, which can be an especially useful strategy when dealing with huge volumes of data in the data mining context. In evaluating the performance of a store, typically annual data are used. To obtain more objective and informative comparison, it is useful to employ multi-years of annual data.

In terms of *product life-cycle,* it is also quite important to study the popularity trend of a product. Such a trend could be different from region to region for the same product. By understanding the popularity trends of the available products, corporate management can take action to further promote a popular product, or revamp/delete a declining product. For such a study to be meaningful, many years of annual data may be needed.

For time series analysis using annual data, it is unlikely that enough data points will be available for conducting typical ARIMA modeling. When limited data are available, graphical display and careful analysis of each time series are crucial for reaching correct conclusions.

4.4 Some Data Warehousing Issues at the Corporate Level

As discussed above, time series data mining at the corporate level may employ daily, weekly, monthly, quarterly, or annual data depending upon the application. With the large number of series potentially of interest and the number of stores involved, data warehousing at the corporate level requires careful consideration. While one approach is to consider a centralized data warehouse integrating data from various stores, alternatively operational data can be maintained locally as "data marts" which in turn feed the corporate data warehouse. The latter approach may be preferred since the store level analyses may need all local data, whereas the analyses at the corporate level may only require a subset of the time series generated at the store level.

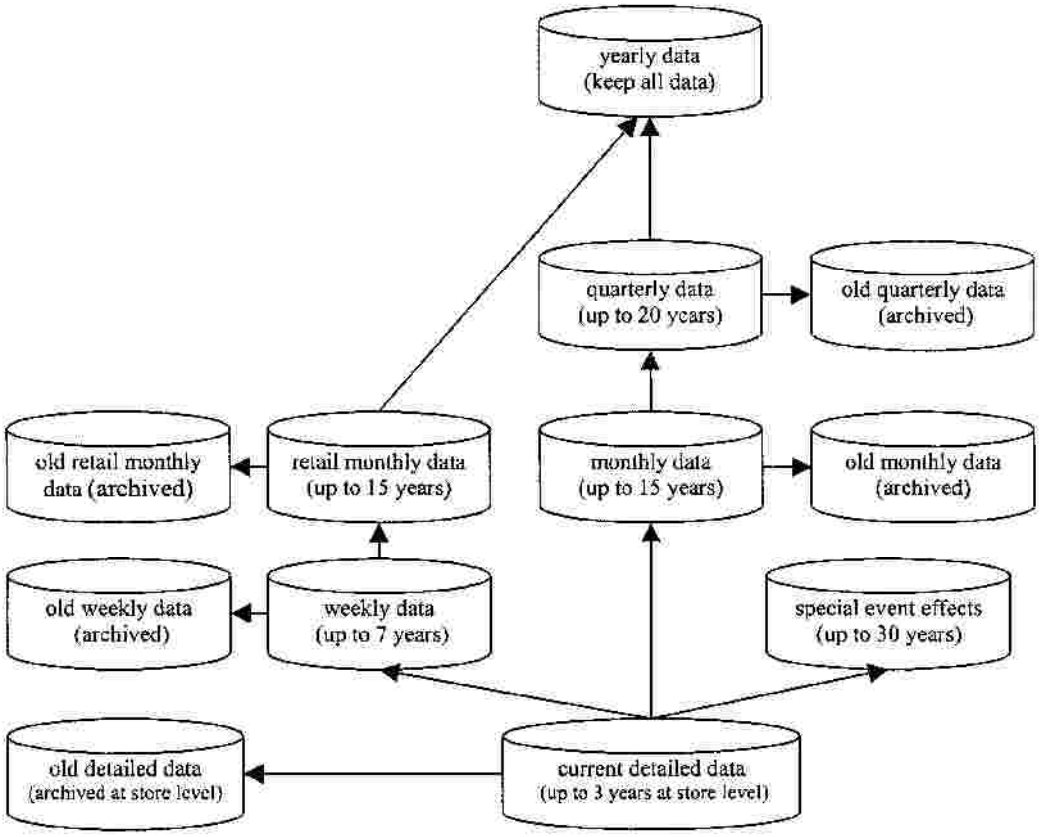
In addition to the issues raised in the beginning of this section, it is important to note that depending upon the application and the granularity of the time series, a certain minimum length of time series is needed in order to develop an adequate model and generate forecasts with acceptable accuracy. For daily time series, a few years (say 2-3 years or more) of data will be sufficient to begin time series modeling and forecasting. For weekly time series, five years or longer may be needed. For monthly and quarterly time series, 10 years or longer would be ideal. For annual data, it is difficult to perform a straightforward univariate ARIMA modeling, and in this case, the longer the series the better. Such disparate requirements of data length can be suitably met by using the hierarchical organization of dimensioned data that is often employed in data warehouses (Chaudhari and Dayal, 1997). For example, the demand data can be organized along the dimensions of Store, Product, and Time and these dimensions then can have hierarchies defined; for example, Product can be organized along product-type, category etc., and Time can define a hierarchy of day, week, month, etc.

As mentioned, store level analyses may employ aggregated time series based on quarter-hour data. Considering this as the lowest granularity, a hierarchy on the time dimension can provide a database organization as shown in Figure 4. Here, the current detailed database holds quarter-hour data up to a 3-year history at the store level, allowing sufficient data points for intra-day and inter-day analyses required at each store. The next level of aggregation can hold weekly, monthly and quarterly data, and similarly weekly data may be aggregated into retail-month and annual data, *etc.*, with higher levels of aggregation corresponding to time series being maintained over longer histories. Note that special event effects, for example effects due to Independence Day, may also be stored separately, facilitating direct analyses that would otherwise necessitate costly retrievals involving archived historical data.

The data at different levels of aggregation may be pre-computed and stored to facilitate quick query response. Alternatively, to economize on storage, part of the data can be computed at query time; here, the archival policy will need to establish that data for the defined aggregate levels be computed from the more detailed data before they are retired from the data warehouse.

The data warehouse can store data as depicted in Figure 4 for different stores and products; considering a typical star schema, these can form other dimensions. Variations of a simple star schema should, however, be considered and can provide a more efficient design. Further, specialized methods for handling time series data can be useful in this context. Relational databases do not naturally store data in time order, leading to cumbersome access. With specialized time series databases, a time series can be stored as an object -- one single logical vector, with the same database record holding all the data pertaining to the time series. Here, adding data to a time series involves appending it to the record instead to adding another independent record as in a regular relational table; a complete time series can thus be readily accessed. New object-relational databases also provide direct support for time series as customizable complex data objects together with specialized operations for their manipulation and analysis, which can be useful in this context.

Figure 4. Aggregations in the corporate data Warehouse



5. Summary and Discussion

Data mining is an emerging discipline that is used to extract information from large databases. A substantial amount of work in this area has focused on cross-sectional data. In this paper we have presented an approach on time series data mining in which automatic time series model identification and automatic outlier detection and adjustment procedures are employed. Although modern business operations regularly generate a large amount of data, we have found very little published work that links data mining with time series modeling and forecasting applications. By using automatic procedures, we can easily obtain appropriate models for a time series and gain increased knowledge regarding the homogenous patterns of a time series as well as anomalous behavior associated with known and unknown events. Both types of knowledge are useful for forecasting a time series. The use of automatic procedures also allows us to handle modeling and forecasting of a large number of time series in an efficient manner.

The time series data mining procedures discussed in this paper have been implemented in a fast-food restaurant franchise. It is easy to see that similar approach can be applied to other business operations and reap the benefits of time series data mining. More generally, an interesting review article on the current and potential role of statistics and statistical thinking to improve corporate and organizational performance can be found in Dransfield, Fisher and Vogel (1999).

In this paper, we employ univariate ARIMA models for time series data mining. The concept can be extended to multi-variable models such as multiple-input transfer function models, and multivariate ARIMA models. The former can be viewed as an extension of multiple regression models for time series data, and the latter is an extension of univariate ARIMA models. For univariate time series modeling, we may also consider certain classes of non-linear and non-parametric models if it is deemed to be more appropriate for the application.

Acknowledgements

The authors would like to thank Jason Fei for his assistance on the data analysis in this paper. This research was supported in part by grants from The Center for Research in Information Management (CRIM) of the University of Illinois at Chicago, and Scientific Computing Associates Corp.

REFERENCES

Box, G.E.P. and Jenkins, G.H. (1970). Time Series Analysis: Forecasting and Control. San Francisco: Holden Day. (Revised edition, 1976.)

Box, G.E.P. and Tiao, G.C. (1975). "Intervention Analysis with Application to Economic and Environmental Problems". Journal of the American Statistical Association 70: 70-79.

Chang, I., Tiao, G.C. and Chen, C. (1988). "Estimation of Time Series Parameters in the Presence of Outliers". Technometrics 30: 193-204.

Chaudhuri, S. and Dayal, U. (1997). "An Overview of Data Warehousing and OLAP Technology". ACM SIGMOD Record 26(1), March 1997.

Chen, C. and Liu, L.-M. (1993a). "Joint Estimation of Model Parameters and Outlier Effects in Time Series." Journal of the American Statistical Association 88:284-297.

Chen, C. and Liu, L.-M. (1993b). "Forecasting Time Series with Outliers." Journal of Forecasting 12:13-35.

Dransfield, S.B., Fisher, N.I., and Vogel, N.J. (1999). "Using Statistics and Statistical Thinking to Improve Organisational Performance." International Statistical Review 67**:** 99-150 (with Discussion and Response).

Fayyad, U. M. (1997). "Editorial." Data Mining and Knowledge Discovery 1: 5-10.

Fox, A.J.(1972). "Outliers in Time Series". Journal of the Royal Statistical Society, Series B 34: 350-363.

Friedman, J. H. (1997). "Data Mining and Statistics: What's the Connection ?" Proceedings of Computer Science and Statistics: the 29th Symposium on the Interface.

Glymour, C., Madigan, D, Pregibon, D. and Smyth, P. (1997). "Statistical Themes and Lessons for Data Mining." Data Mining and Knowledge Discovery 1: 11-28.

Hand, D.J. (1998). "Data Mining: Statistics and More?" The American Statistician 52:112-118.

Hillmer, S.C. and Tiao, G.C. "Likelihood Function of Stationary Multiple Autoregressive Moving Average Models." Journal of the American Statistical Association 74: 652-660.

Liu, L-M. (1980). "Analysis of Time Series with Calendar Effect." Management Science 26: 106-112.

Liu, L-M. (1986). "Identification of Time Series Models in the Presence of Calendar Variation." International Journal of Forecasting 2: 357-372.

Liu, L.-M. (1989). "Identification of Seasonal ARIMA Models Using a Filtering Method." Communication in Statistics A18: 2279-2288.

Liu, L.-M. and Lin, M.-W. (1991). "Forecasting Residential Consumption of Natural Gas Using Monthly and Quarterly Time Series." International Journal of Forecasting 7: 3-16.

Liu, L-M. and Chen, C. (1991) "Recent Developments of Time Series Analysis in Environmental Impact Studies." Journal of Environmental Science and Health A26: 1217-1252.

Liu, L.-M. and Hudak, G.B. (1992). Forecasting and Time Series Analysis Using the SCA Statistical System: Volume 1. Chicago: Scientific Computing Associates Corp.

Liu, L.-M. (1993). "A New Expert System for Time Series Modeling and Forecasting." Proceeding of the American Statistical Association - Business and Economic Section 1993: 424-429.

Liu, L.-M. (1999). Forecasting and Time Series Analysis Using the SCA Statistical System: Volume 2 . Chicago: Scientific Computing Associates Corp.

Reilly, D.P. (1980). "Experiences with an Automatic Box-Jenkins Modeling Algorithm." Time Series Analysis – Proceedings of Houston Meeting on Time Series Analysis. North Holland Publishing.

Reynolds, S.B., Mellichamp, J.M., and Smith, R.E. (1995). "Box-Jenkins Forecast Model Identification." AI Expert, June 1995: 15-28.

Thompson, H.E. and Tiao, G.C. (1971). "Analysis of Telephone Data: A Case Study of Forecasting Seasonal Time Series." The Bell Journal of Economics and Management Science 2: 515-541.

Tsay, R.S. and Tiao, G.C. (1984). "Consistent Estimates of Autoregressive Parameters and Extended Sample Autocorrelation Function for Stationary and Non-stationary ARMA Models". Journal of the American Statistical Association 79: 84-96.

Tsay, R.S. and Tiao, G.C. (1985). "Use of Canonical Analysis in Time Series Model Identification." Biometrika 72: 299-315.

Tsay, R.S. (1988). "Outliers, Level Shifts, and Variance Changes in Time Series." Journal of Forcasting 7: 1-20.

Weiss, S. M. and Indurkhya, N. (1998). Predictive Data Mining. San Francisco: Morgan Kaufmann Publishers.

Widom, J. (1995). "Research Problems in Data Warehousing". Proceedings of 4th International Conference on Information and Knowledge Management (CIKM), November 1995.