Analysis of call centre arrival data using singular value decomposition

Haipeng Shen[1],*,[†] and Jianhua Z. Huang[2],‡

[1] *Department of Statistics and Operations Research, University of North Carolina at Chapel Hill, U.S.A.*
[2] *Department of Statistics, Texas A&M University, U.S.A.*

SUMMARY

We consider the general problem of analysing and modelling call centre arrival data. A method is described for analysing such data using singular value decomposition (SVD). We illustrate that the outcome from the SVD can be used for data visualization, detection of anomalies (outliers), and extraction of significant features from noisy data. The SVD can also be employed as a data reduction tool. Its application usually results in a parsimonious representation of the original data without losing much information. We describe how one can use the reduced data for some further, more formal statistical analysis. For example, a short-term forecasting model for call volumes is developed, which is multiplicative with a time series component that depends on day of the week. We report empirical results from applying the proposed method to some real data collected at a call centre of a large-scale U.S. financial organization. Some issues about forecasting call volumes are also discussed. Copyright © 2005 John Wiley & Sons, Ltd.

KEY WORDS: anomaly detection; call centre; data reduction; feature extraction; forecasting call volume; singular value decomposition

1. INTRODUCTION

Call centres are modern service networks in which customer service representatives (CSRs) provide services to customers via telephones. They have become a primary contact point between service providers and their customers. For example, Reference [1] estimates that more than 70% of all customer–business interactions are handled in call centres in 2002; Reference [2] states that the call centre industry in U.S. employs more than 3.5 million people, or 2.6% of its workforce. Thus managing call centre operations efficiently is playing a more and more important role in our modern business world.

Well-run call centres seek to balance CSR utilization and service quality according to some pre-specified measure, and to achieve that, call centre managers use queueing-theoretic models

*Correspondence to: H. Shen, Department of Statistics and Operations Research, University of North Carolina at Chapel Hill, Chapel Hill, NC 27599, U.S.A.
[†] E-mail: haipeng@email.unc.edu
[‡] E-mail: jianhua@stat.tamu.edu

such as $M/M/N$ queue or Erlang-C to manage call centre operations. Inputs to these queueing models are statistics concerning system primitives such as call arrival rates, service durations, customer abandonment (patience) behaviour and even retrials. These statistics can be obtained through statistical analysis of call centre *operational* data, which record various call centre operations and reflect the physical process by which calls are handled. As pointed out in Reference [3], 'the modelling and control of call centres must necessarily start with careful data analysis'. Furthermore, statistical analysis can be used to validate and calibrate queueing-theoretic models. Reference [4] provides a thorough statistical analysis of operational data from a small Israeli bank call centre, where several common queueing model assumptions are rejected empirically; however some queueing-theoretic results are shown to be robust against such violations while a few others are not.

On the other hand, call centres are increasing quickly both in the size and the complexity of their operations. During its routine operation, a medium-to-large call centre could generate a vast amount of data. For example, a call centre with several hundred CSRs can handle 300K or so calls every day, which could result in a couple of gigabytes' worth of call-by-call data each month. The data-rich environment suggests that call centres have a lot of potential for statistical analysis.

However, despite the importance of statistics and the tremendous amount of data available, the literature on statistical inference in the world of call centres is still sparse and there is an urgent and growing need for serious research efforts along this line. Reference [5] provides a comprehensive bibliography of 250-plus call-centre-related research papers in a wide range of disciplines; only 17 papers are listed on the statistics and forecasting of call centre data. A big gap exists between the practice of statistics and the prevalent needs in call centre modelling. In addition, the large volume of data is also a challenge for existing statistical techniques, which are traditionally developed for small-to-medium data sets.

The first component of call centre operational data is call arrival data, which records the times at which calls arrive to the centre. Call-arrival data is essential for understanding the first system primitive—call arrivals, as well as developing call volume forecasting models.

Forecasting call volumes is an important issue for call centre operations. For highly utilized call centres, accurate forecasts are necessary for workload forecasting, CSR allocation and capacity planning. There exist some regularities within call-arrival data, such as inter-day dependence and intra-day dependence. By inter-day dependence, we mean correlation between volumes of consecutive days and possible weekly, monthly, seasonal and even yearly cycles in call volumes. Intra-day dependence refers to correlation between arrivals in different time periods (morning/afternoon/night) within the same day. Good forecasting models should capture at least one dependence structure or even both. See References [4, 6] for forecasting models which make use of these two dependence structures, respectively.

In addition to these aforementioned regularities, there may also exist various kinds of anomalies in the data as shown later in Section 5.1. For example, call-arrival patterns can be very unusual due to holidays and business campaigns. Additionally, the data is typically collected by systems like *interactive voice response unit* (VRU) and *automatic call distributor* (ACD). System (hardware/software-related) failures could lead to anomalies in the data as well. These anomalies usually have very different arrival characteristics from regular operations, which may weaken or even destroy the correlation structures in the data. So one should try to identify them first, and model them separately from the regular arrival patterns. Call centre managers often have a good sense of the first type of

anomalies through past experience and business marketing information. The second type of anomalies, however, cannot be foreseen and is harder to identify. Researchers and model developers usually analyse historical data which may consist of both. One problem with massive data is that researchers cannot look at the whole data with one snapshot. Thus, it is desirable to have an automatic procedure that is effective for identifying these hidden anomalies all together.

The authors hope to contribute to the literature by introducing a statistical method—singular value decomposition (SVD), which can be used to effectively analyse (large) call centre data (arrival data in particular) and to provide insights for model building. The SVD method appears to be a simple tool to automatically identify anomalies in a unified way. The method is useful to extract significant features from the data, such as day-of-week effect, average daily arrival pattern and operating hour switching. Moreover, a high degree of data reduction can be achieved by using a few significant features to summarize the original data. The reduced data can be effectively used in descriptive and formal statistical analysis. Our method is illustrated via an application of call arrival data gathered at an U.S. inbound call centre in 2002. For this real data example, the SVD suggests that total daily call volume and average daily arrival pattern are the two key components. Furthermore, a call volume forecasting model is built using these two components, which is multiplicative and has an auto-regressive (AR) time series structure that depends on day-of-week.

This paper is structured as follows. Section 2 provides a brief description of the call centre arrival data which is used as an application of our methodology. SVD is then reviewed in Section 3. Section 4 presents general techniques based on SVD that one can use to analyse call-arrival data. A case study is presented in Section 5. We conclude in Section 6 and discuss some other potential applications of the methodology.

2. CALL CENTRE ARRIVAL DATA

The data motivating our study was gathered at an inbound call centre of a major northeastern U.S. financial firm in 2002. The original database has detailed information about every call that got connected to this call centre during each day of the year (except 6 weekdays where the data collecting equipment went out of order). The centre opens normally from 7 AM to mid-night. For the current study, we are interested in understanding the arrival pattern of calls to the service queue. As a result, the portion of the data of interest to us here involves the information about the time every call arrives to the service queue.

For a particular day, we divided the 17-h operating period into 408 150-s intervals, and recorded the number of calls arriving to the service queue during each interval. This data aggregation procedure, which introduces smoothing to the raw data, is commonly used in the literature when estimating call arrival rate; see References [4, 7]. The data we obtain is a 360×408 count matrix, with each row corresponding to a particular day in 2002 and each column corresponding to one specific 150-s interval between 7 AM and mid-night. There are 257 weekdays and 103 weekends in the data. Here the length of the intervals, 150 s, is chosen rather subjectively for illustration purpose only. One nice thing about having a call-by-call database is that one can easily derive arrival counts for intervals of arbitrary length. If one decides to use 15-min or $\frac{1}{2}$-h intervals, it will result in further smoothing of the raw data.

3. SVD

This section provides a brief review of SVD that is well-known in matrix algebra. Let $X = (x_{ij})$ denote an $n \times m$ matrix of real-valued data with a rank $r \leqslant \min(n, m)$. In the case of call arrival data from a call centre, x_{ij} can be the number of incoming calls during the jth time period of the ith day. The elements of the ith row of X form an m-dimensional vector $\mathbf{r}_i^t$, which we refer to as the *intra-day call volume profile* of the ith day. Similarly, the elements of the jth column of X form an n-dimensional vector $\mathbf{c}_j$, which we refer to as the *inter-day call volume record* of the jth time period.

The SVD of the data matrix X can be expressed as follows:

$$X = USV^t \tag{1}$$

where U is an $n \times m$ matrix with orthonormal columns, S is an $m \times m$ diagonal matrix, and V^t is an $m \times m$ orthogonal matrix. The columns of U, $\{\mathbf{u}_k\}$, are the left singular vectors, and form an orthonormal basis for the inter-day call volume records $\mathbf{c}_j$'s, so that $\mathbf{u}_i^t \mathbf{u}_j = 1$ for $i = j$, and $\mathbf{u}_i^t \mathbf{u}_j = 0$ otherwise. We name $\{\mathbf{u}_k\}$ the *inter-day feature vectors* in the sense that they summarize features of the inter-day arrival patterns. The rows of V^t (i.e. the columns of V) contain the elements of the right singular vectors, $\{\mathbf{v}_k\}$, and form an orthonormal basis for the intra-day call volume profiles $\mathbf{r}_i^t$'s. The vectors $\{\mathbf{v}_k\}$ are referred to as the *intra-day feature vectors*, which capture the intra-day arrival patterns. The diagonal elements of S are called the singular values, which are usually ordered from high to low. Let $S = \mathrm{diag}(s_1, \ldots, s_m)$. Then $s_1 \geqslant s_2 \geqslant \cdots \geqslant s_r > 0$, and $s_k = 0$ for $r + 1 \leqslant k \leqslant m$. Note that (1) can then be rewritten as

$$X = \sum_{k=1}^{r} s_k \mathbf{u}_k \mathbf{v}_k^t$$

One important result of the SVD of X is the following:

$$X^{(l)} \equiv \sum_{k=1}^{l} s_k \mathbf{u}_k \mathbf{v}_k^t \tag{2}$$

is the closest rank-l matrix approximation to X. Here the term 'closest' simply means that $X^{(l)}$ minimizes the sum of squares of the element-wise differences between X and X^*,

$$\sum_{ij} (x_{ij} - x_{ij}^*)^2$$

for all rank-l matrices $X^* = (x_{ij}^*)$. See Reference [8] for a technical derivation of this result. One consequence of this matrix approximation is that using quantities on the right-hand side of (2), we can have a concise summary of the original data, X. With a small value of l, a substantial amount of data reduction can be achieved. Alternatively, one can think of (2) as a natural way of de-noising or smoothing the original noisy data. The practical issue of choosing an appropriate l is important and will be explained later.

As indicated by the matrix approximation (2), the inter-day call volume records $\mathbf{c}_j$ can be summarized by expressing them as a linear combination of a few inter-day feature vectors $\mathbf{u}_k$, $1 \leqslant k \leqslant l$, and the coefficients in this basis expansion are provided by $s_k v_{kj}$ where v_{kj} is the jth element of $\mathbf{v}_k$. Since l is usually much smaller than n, these coefficients provide a concise summary of the $\mathbf{c}_j$'s which are n-dimensional vectors. Similarly, the intra-day call volume profiles $\mathbf{r}_i^t$ can be summarized by expressing them as a linear combination of the intra-day

feature vectors $\mathbf{v}_k$, $1 \leqslant k \leqslant l$, and the coefficients in this basis expansion are provided by $s_k u_{ki}$, where u_{ki} is the ith element of $\mathbf{u}_k$. Again these coefficients provide a concise summary of the $\mathbf{r}_i^t$'s which are m-dimensional vectors. As discussed below in Section 4, these coefficients are good candidates for data visualization and statistical modelling.

4. SVD ANALYSIS OF CALL CENTRE DATA

In this section, we discuss several SVD-related statistical techniques, which can be used to analyse call centre data. These techniques are illustrated through a case study in Section 5.

The diagonal values of S (i.e. s_k) make up the *singular value spectrum*. The quantity $s_k^2/(\sum_i s_i^2)$ shows the relative importance of the kth singular vector pairs in explaining the data, and we refer to it as the kth 'energy percentage'. These percentages can be plotted versus k, and the resulting plot is usually called a *scree plot*. One can decide on the number of significant components, l, using the plot. If the original variables are linear combinations of l underlying variables, in addition to some low-level noise, the plot will tend to drop sharply for the first l singular values associated with those underlying variables and then much more slowly for the remaining singular values. In practice, one tries to find such an 'elbow' so that the singular values plotted to its right are ignored because they are assumed to be mainly due to noise.

The inter-day feature vectors $\{\mathbf{u}_k\}$ provide an orthonormal basis for the inter-day call volume records $\{\mathbf{c}_j\}$. The first few of them, associated with the significant singular values, can be individually plotted over day as a time series. These time series plots may reveal inter-day dependence and interesting inter-day call arrival patterns. The vectors $\{\mathbf{u}_k\}$ also give a representation of the intra-day profiles using the intra-day feature vectors. More precisely, the projection of the intra-day profile $\mathbf{r}_i$ on the intra-day feature vector $\mathbf{v}_k$ has a projection coefficient $q_{ik} = \mathbf{r}_i^t \mathbf{v}_k$, which equals $s_k u_{ki}$ since q_{ik} is the (i, k)th element of $XV = US$. Thus, a scatter plot of $\{(q_{ik}, q_{ik'}), i = 1, \ldots, n\}$ for a fixed $\{k, k'\}$ pair, or equivalently, a scatter plot of $\{(u_{ki}, u_{k'i}), i = 1, \ldots, n\}$ (note that s_k and $s_{k'}$ do not change with i), can be used to find differences among different days. Day-of-week effect, holiday effect and other anomalies can be detected by such plots. The projection coefficients $\{q_{ik}\}$ or $\{u_{ki}\}$ can also be used for analysis of variance (ANOVA), cluster analysis, regression analysis, time series analysis, and so on. For example, one might build an ANOVA model on $\{u_{ki}\}$ for a fixed k using day-of-week as one factor. Another example is given in Section 5.3 where a time series model is built upon $\{u_{1i}\}$ as part of a volume forecasting model.

Similarly, the intra-day feature vectors $\{\mathbf{v}_k\}$ provide an orthonormal basis for the intra-day call volume profiles $\{\mathbf{r}_i\}$. The first few significant vectors, can be separately plotted over time periods within a day and searched for recognizable intra-day call arrival patterns. The projection of the inter-day call volume record $\mathbf{c}_j$ on the inter-day feature vector $\mathbf{u}_k$ has a projection coefficient $p_{jk} = \mathbf{u}_k^t \mathbf{c}_j \ (= s_k v_{kj})$, which is the (k, j)th element of $U^t X \ (= SV^t)$. A scatter plot of $\{(p_{jk}, p_{jk'}), j = 1, \ldots, m\}$ (or $\{(v_{kj}, v_{k'j}), j = 1, \ldots, m\}$) for a fixed $\{k, k'\}$ pair can be used to find varying call arrival patterns during a day.

5. CASE STUDY

In this section, we would like to apply the general SVD analysis techniques described in Section 4 to the call centre data mentioned in Section 2. A look at the 'energy percentage', $s_k^2/(\sum_i s_i^2)$,

suggests that the first singular vector pair explains 98.7% of the total energy while the second pair explains about 40% of the remaining energy. We focus on the first two pairs of feature vectors in the following analysis.

Figure 1 is a scatter plot of the first two inter-day feature vectors, $\mathbf{u}_1$ and $\mathbf{u}_2$. To identify the day-of-week effect, the points are coded using different colours and symbols according to their corresponding days of the week as shown in the legend. There are approximately six clusters in Figure 1. From right to left, the first cluster, and also the largest, consists of points from the five weekdays with about 12 points scattered around the main cluster. These 12 points are 'suspected' anomalies and will be looked at in detail in Section 5.1. Then there are two Saturday clusters and two Sunday clusters, respectively. If one looks at the top Sunday cluster closely, one can identify a Tuesday point hidden inside the cluster, which is identified to be a holiday later. The last cluster is at the left end of the plot, which consists of two Mondays, one Wednesday and two Thursdays. These five days are also identified to be holidays later. In total we have identified 18 'suspected' weekday anomalies, which are indicated using arrows in Figure 1. These 18 days are further divided into three categories based on why they appear to be anomalies in Section 5.1.

Figure 1 shows that the first two inter-day feature vectors can separate out the weekdays, the Saturdays and the Sundays to a great extent. There is also much more variability among the weekdays than the weekends, which suggests one to look at the weekdays and the weekends separately. Below, in Section 5.1, SVD is applied to the weekday data to identify anomalies and categorize them. The weekend data is used in Section 5.2 to show how one can detect significant patterns using SVD. One puzzle that we will solve is why there are two clusters rather than one for the Saturdays and also for the Sundays. The weekday data (without the anomalies) is also analysed in Section 5.2 to extract regular features among weekdays. Furthermore, the outputs from the SVD are used to build a forecasting model for call volumes in Section 5.3.
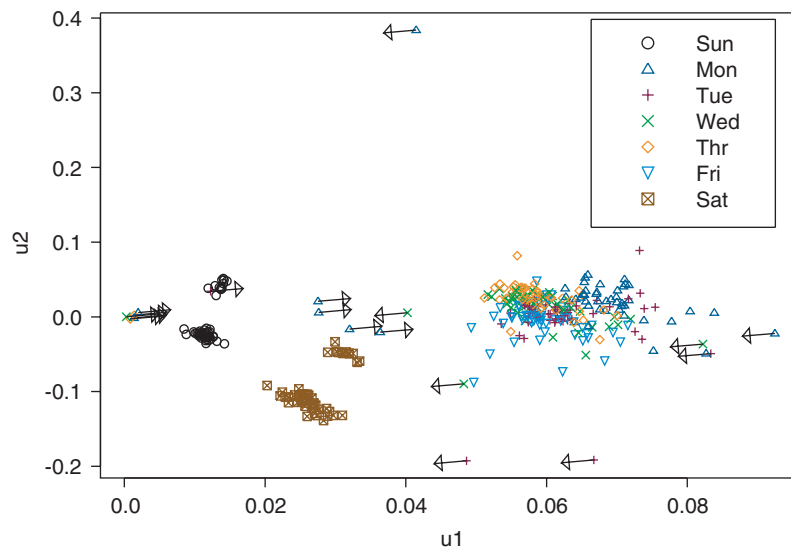

Figure 1. Scatter plot of $\mathbf{u}_1$ and $\mathbf{u}_2$ for all days.

Appl. Stochastic Models Bus. Ind., 2005; **21**:251–263

5.1. Anomaly detection

In the current section, we focus on the weekday data and illustrate how one can use SVD, especially the inter-day feature vectors, to identify anomalies. Figure 1 shows that there are 18 anomaly points highlighted separately using arrows. A separate SVD on the weekday only reveals the same set of anomalies. Table I lists the dates of these anomalies and reasons why they are anomalous. As one can see, they are grouped into three categories: holidays, days around holidays and non-holidays with unusual arrival patterns such as system-ill-functioning days.

The ten points indicated by the right arrows correspond to ten major observed U.S. holidays in 2002. It's very reasonable and well accepted by call centre managers that call volumes during major holidays are usually very low. In practice call centre managers will run special staffing operations during these days. The SVD confirms this common perception automatically, which in turn suggests the usefulness of this particular analysing tool.

Indicated by the left arrows, the remaining eight outliers seem to fall into two groups as shown in Table I. The first group includes those days which are before or after major holidays such as Thanksgiving, Christmas and New Year. In general, call volumes will be higher than usual after major holidays and people also start to call earlier. For example, the number of arrivals per interval for Jan. 2 is plotted in the left panel of Figure 2, along with the other Wednesdays in Jan. 2002. The interval volumes in Jan. 2 seem to be uniformly larger than the other Wednesdays, which is rather easy to understand given that Jan. 2 is the first 'regular' business day of the year and also after the long holiday season; the volume also seems to increase faster in the morning and drop faster in the afternoon. Dec. 2 and 3 fall into the same category because they are the first two weekdays after the long Thanksgiving weekend and the start of the holiday shopping season. On the other hand, for business days right before major holidays such as Dec. 24 and 31, call volumes are usually lower and people stop calling the bank earlier than usual.

The second group includes those days with possible system errors and other abnormal behaviours. The call volume during Apr. 1 is plotted in the right panel of Figure 2. There are almost no calls connected to the service queue between 7:15 and 11:45 AM, indicated by the two vertical dashed purple lines. It turns out that the data-recording device malfunctioned during this period and failed to record any data. One could say that the system played a joke on the call

Table I. Eighteen anomalies in 2002.

| Date	Reason	Date	Reason
Holiday		*Holiday related*	
Jan. 1	New Year's day	Jan. 2	After New Year's day
Jan. 21	MLK day	Dec. 2/3	After Thanksgiving
Feb. 18	Washington's birthday	Dec. 24	Christmas eve
May. 27	Memorial day	Dec. 31	New Year's eve
Jul. 4	Independence day		
Sep. 2	Labor day	*System related/other*	
Oct. 14	Columbus day	Apr. 1	System error
Nov. 11	Veterans' day	Jul. 24	System error
Nov. 28	Thanksgiving day	Dec. 24	System error
Dec. 25	Christmas day	Sep. 11	'Emotional' day
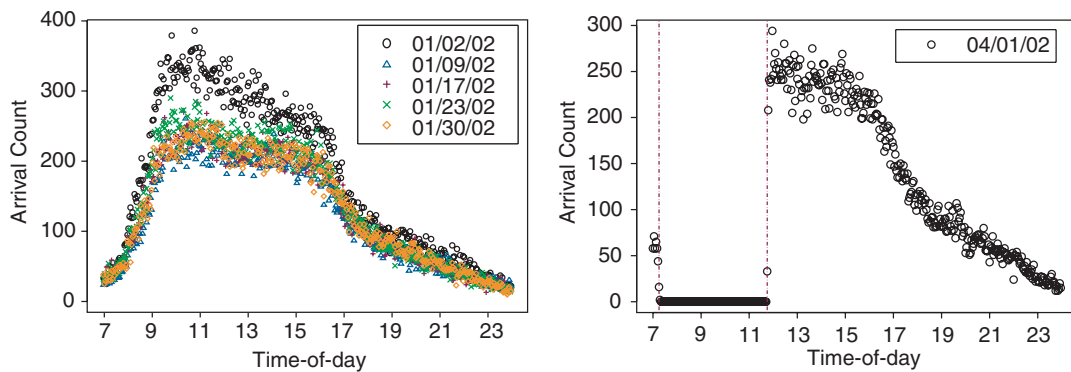
Figure 2. Arrival patterns for two anomalous days in 2002.

centre manager on this April Fool's Day. Jul. 24 and Dec. 24 (*not shown here*) also belong to this group. Another day in the second group is Sep. 11, whose call arrival pattern is very normal except that the interval volumes are uniformly smaller. This might be due to the special meaning of this 'emotional' day after the terrorist attack on Sep. 11, 2001. A lot of memorial services are held on this day.

In summary, SVD appears to be a useful tool to automatically detect anomalies. Note that the collection of anomalies we provided here is not meant to be complete but rather the most significant ones. If one wants, SVD can help to identify some other less significant anomalies very easily. Our purpose is to illustrate how one can use SVD to achieve that goal.

5.2. Feature extraction

In this section, we show how SVD can be used to automatically extract significant features hidden in the data. We first look at the weekend data.

Figure 1 shows that there are two Saturday clusters and two Sunday clusters. The two panels of Figure 3 show the time series plots of the first two inter-day feature vectors, $\mathbf{u}_1$ and $\mathbf{u}_2$, respectively. As one can see, $\mathbf{u}_1$ distinguishes Saturdays from Sundays while $\mathbf{u}_2$ contains information about changes within Saturdays and Sundays. It is clear that $\mathbf{u}_2$ increases significantly for both Saturdays and Sundays since March 30 and March 31.

A follow-up communication with the call centre management reveals one operational change. Prior to March 30, the call centre operates between 7 AM and mid-night during weekends; since then, the weekend operating hour reduces to 8 AM–5 PM. It is clear that $\mathbf{u}_2$ captures exactly this working hour switch. This shows that SVD is very powerful in using very few singular vectors to extract significant patterns from the data, and the whole process can be automated.

As discussed earlier in Section 3, the intra-day feature vectors summarize information related to the different intervals within each day. Figure 4 plots the first two intra-day feature vectors, $\mathbf{v}_1$ and $\mathbf{v}_2$. It turns out that $\mathbf{v}_1$ shows the overall average daily arrival pattern among the weekends, which has a single mode before noon. The average pattern changes sharply around 8 AM and
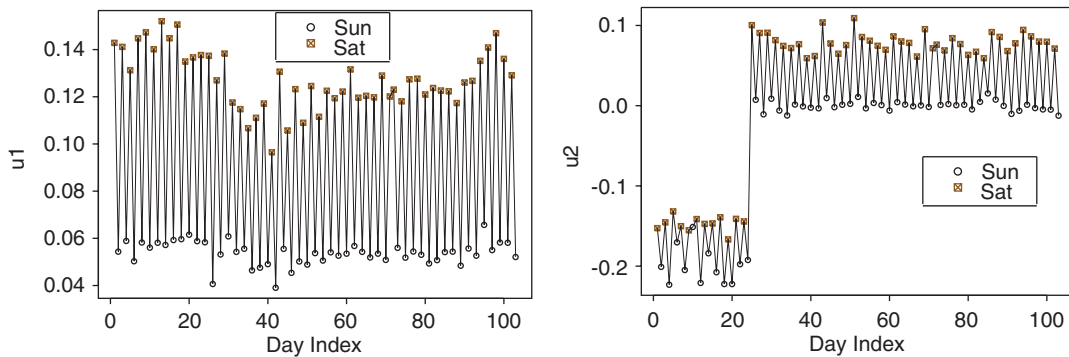
Figure 3. Time series plots of $\mathbf{u}_1$ and $\mathbf{u}_2$ for weekends.
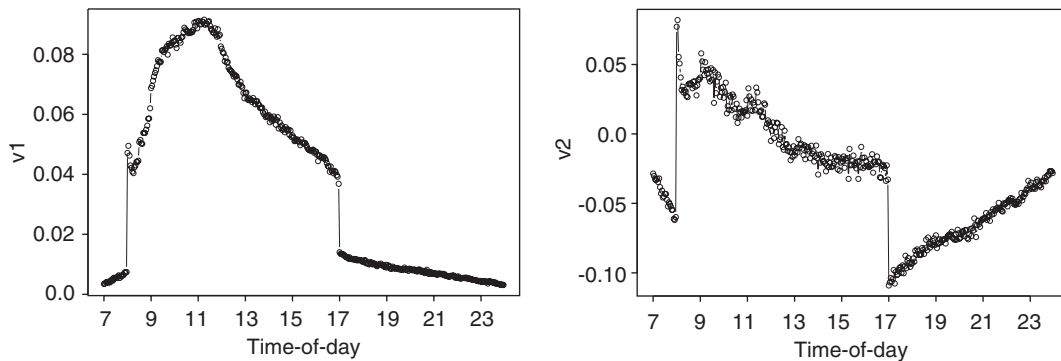
Figure 4. Plots of $\mathbf{v}_1$ and $\mathbf{v}_2$ for weekends.

5 PM, which are caused by the working hour switch. We also find that $\mathbf{v}_2$ summarizes the difference between the overall average arrival pattern and the average arrival pattern before the working hour reduction.

As for the weekday data, the 18 anomalies identified in Section 5.1 are excluded before applying SVD to the remaining count matrix. Figure 5 plots the first two intra-day feature vectors, $\mathbf{v}_1$ and $\mathbf{v}_2$. Similar to Figure 4, $\mathbf{v}_1$ summarizes the average arrival pattern during a regular weekday. The pattern is bimodal with one morning mode, one afternoon mode and a lunch-break dip in the middle, which is clearly different from the weekend arrival pattern. Note that there is also an early-morning dip, which means some customers contact the centre as soon as it opens.

There is no working hour change among the weekdays as in the weekends. As a result, the second intra-day feature vector $\mathbf{v}_2$ for the weekdays has a different meaning. Considered together with $\mathbf{u}_2$, $\mathbf{v}_2$ can be shown to capture a specific feature that Fridays usually have above average call volumes in the morning and below average volumes in the afternoon with the opposite is true for Mondays. Note that this is a second-order effect comparing to the average daily arrival pattern.

To summarize, SVD can easily separate the weekends from the weekdays, and even distinguish the Saturdays from the Sundays. The weekdays seem to form three groups as well,
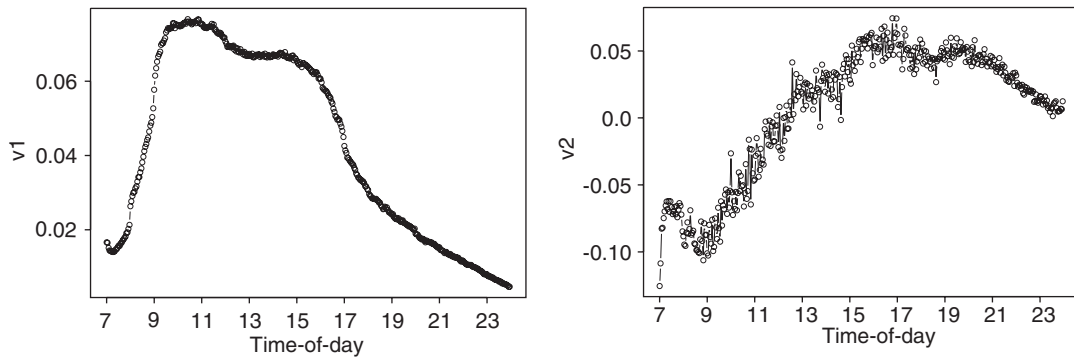
Figure 5. Plots of $\mathbf{v}_1$ and $\mathbf{v}_2$ for weekdays.

Mondays, Fridays and the remaining weekdays. SVD can also extract features like average daily arrival patterns and working hour switch automatically.

5.3. Forecasting call centre arrivals

We propose the following modelling procedure based on SVD. One should first apply SVD to detect possible anomalies and exclude them prior to any formal modelling effort. The existence of anomalies can sometimes mask or weaken significant features in the data that are useful for modelling purpose. Then one should repeat SVD with the cleaned data and decide on the number of significant singular values. The resulting significant singular vectors then can be explored to extract features and eventually modelled formally.

In the current section, we use the results obtained from the SVD analysis, and develop a forecasting model for call arrivals. As shown in Section 5.2 arrival patterns are different during weekdays and weekends. Here we will focus on the weekday data without those anomalies.

The cleaned data has 239 regular weekdays. The first pair of singular vectors clearly dominates the other pairs since it explains about 99% of the total energy. Let x_{ij} denote the number of calls during the jth interval of the ith day. Using the first pair of singular vectors to approximate the original count matrix, $X = (x_{ij})$, we write

$$x_{ij} = s_1 u_{1i} v_{1j} + \varepsilon_{ij} \tag{3}$$

where u_{1i} is the value of $\mathbf{u}_1$ at day i, v_{1j} is the value of $\mathbf{v}_1$ at the jth interval, and ε_{ij} is the approximation error.

We first discuss three findings about $\mathbf{u}_1$, which will naturally suggest a forecasting model for x_{ij}. First, $\mathbf{u}_1$ is highly correlated with total daily call volume with a correlation of 0.997. The nearly perfect linear relationship suggests that prediction of daily call volume is almost equivalent to prediction of $\mathbf{u}_1$. Actually $\mathbf{u}_1$ has a slightly larger prediction power than the daily call volume because $\mathbf{u}_1$ is less noisy than the daily volume.

Secondly, $\mathbf{u}_1$ for Mondays can be used to predict the rest of the week. There exist very strong positive linear relationships between u_1 of Mondays and the other weekdays with the correlations being 0.916, 0.877, 0.825 and 0.755, respectively. The correlation is 0.887 between Mondays' u_1 and the sum of u_1's from the rest of the weekdays in the same week. This is very useful for medium-term volume forecasting and capacity planning. The strong relationships will

be diminished if those anomalies identified earlier are included. This illustrates the importance of detecting and excluding anomalies before any further formal modelling.

Thirdly, $\mathbf{u}_1$ of two consecutive weekdays are highly correlated, and even more correlated conditioning on day-of-week. Figure 6 plots the current day's u_1 against the next weekday's u_1. Different colours and symbols are used according to the current day's day-of-week. There is a positive linear relationship with a correlation of 0.51. The Mon/Tue, Tue/Wed, Wed/Thu, Thu/Fri and Fri/Mon pairs seem to lie along several lines with similar slopes but different intercepts. The individual correlations are 0.916, 0.945, 0.904, 0.888 and 0.817, respectively. The pooled correlation increases to 0.758 after leaving out the Fri/Mon pairs, which has the weakest relationship. This is easy to understand given that the Fri/Mon pairs are actually 2 day apart.

The aforementioned findings about $\mathbf{u}_1$ suggest the following varying-coefficient AR(1) model for $\mathbf{u}_1$,

$$u_{1i} = a(d_{i-1}) + b(d_{i-1})u_{1(i-1)} + \varepsilon_i \tag{4}$$

where d_i is a factor denoting the day-of-week for the ith day. Figure 6 actually suggests that the slope $b(d_{i-1})$ may be a constant that does not depend on the day-of-week. This motivates us to consider the following common-slope AR(1) model:

$$u_{1i} = a(d_{i-1}) + bu_{1(i-1)} + \varepsilon_i \tag{5}$$

Since model (5) is nested in model (4), one can test formally whether model (5) is sufficient based on the data.

After fitting the two models, model (5) is preferred, because nothing much can be gained by fitting a separate slope for different weekdays as in model (4). The fitted model has a R^2 of 0.83 and a root mean squared error of 0.0025. The residuals appear to be random and normally distributed as suggested by a normal quantile plot. Note that d_i is a factor with five levels
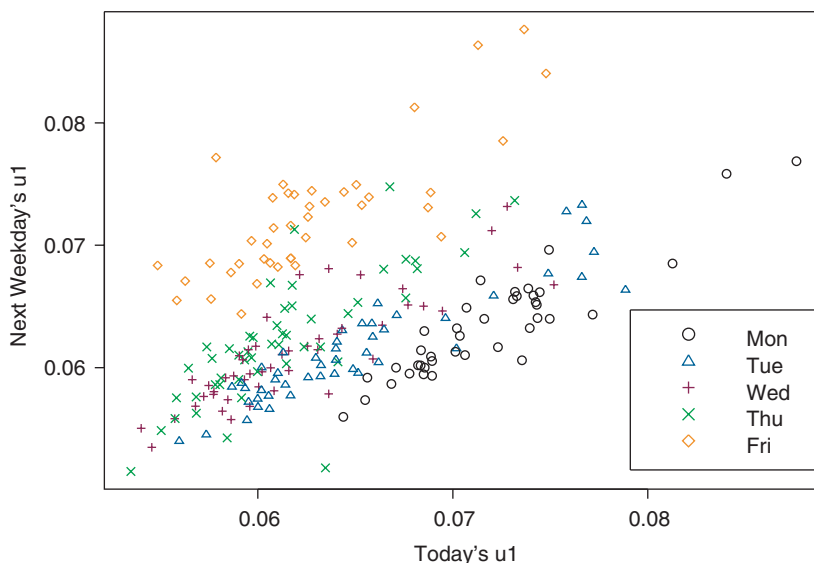

Figure 6. Scatter plot of $\mathbf{u}_1$ between two consecutive weekdays.

corresponding to the 5 weekdays. The fitted model can be written as follows:

$$\hat{u}_{1i} = 0.7892 u_{1(i-1)} + \begin{cases} 0.0067 & \text{if day } i-1 \text{ is a Monday} \\ 0.0102 & \text{if day } i-1 \text{ is a Tuesday} \\ 0.0124 & \text{if day } i-1 \text{ is a Wednesday} \\ 0.0140 & \text{if day } i-1 \text{ is a Thursday} \\ 0.0227 & \text{if day } i-1 \text{ is a Friday} \end{cases} \tag{6}$$

By combining (3) and (5), we obtain the following forecasting model for x_{ij}:

$$x_{ij} = s_1 u_{1i} v_{1j} + \varepsilon_{ij}$$
$$u_{1i} = a(d_{i-1}) + b u_{1(i-1)} + \varepsilon_i \tag{7}$$

where the estimates for s_1, u_{1i} and v_{1j} are derived from the SVD, and $a(d_{i-1})$ and b are given by (6). Given the current day's day-of-week, one can first obtain a forecast for the next weekday's u_1, which can then be multiplied with $\hat{s}_1 \hat{v}_{1j}$ to provide a forecast of the next weekday's call volumes for each time period.

Model (7) is closely related to the multiplicative mixed-effects model proposed by Brown *et al.* [4] for a similar application, where the arrival times are modelled as a time-varying Poisson process and the inter-day dependence is also modelled randomly with an AR(1) component. However, the AR(1) structure in Reference [4] does not depend on day-of-week. The significance of day-of-week effect in our study surely improves the prediction accuracy. Our model is derived via a data analytic approach (namely SVD), which automatically suggests that the features being modelled are the most important ones, which indirectly confirms the model in Reference [4]. In addition the SVD also points out what to model as the second-order effect, i.e. the second singular vector pair, which captures the difference of arrival patterns among different weekdays. In other words, our approach here complements and verifies the more probabilistic approach in Reference [4].

6. DISCUSSION

In this paper, we propose to analyse call centre arrival data using SVD. Although SVD is well-known in matrix algebra, its application to call centre data analysis is new. We illustrate through a case study how SVD can be used as a tool for preliminary data analysis before serious modelling to detect days with unusual arrival patterns, and to extract significant inter-day features such as day-of-week effects and working hour change as well as intra-day features like average arrival patterns. A call volume forecasting model is also developed based on the results of the SVD analysis.

We have been focusing on modelling the first singular vector pair as in the forecasting model (7). The reason is that the first pair explains most (99%) of the energy in the data. There are modelling scenarios where two or more pairs might be needed. For example, the average arrival patterns during weekends are different before and after the working hour change as illustrated above in Section 5.2. Or hypothetically, one can think of a scenario where different weekdays have different arrival patterns. Then more pairs of singular vectors (or intra-day feature vectors) will become significant and therefore should appear in the forecasting model to increase

prediction accuracy. Under the SVD framework, it is straightforward to extend the current model to include additional pairs of feature vectors, and assume the following model:

$$x_{ij} = \sum_{k=1}^{l} s_k u_{ki} v_{kj} + \varepsilon_{ij}$$

$$u_{ki} = f_k(u_{k i_k^*}) + \varepsilon'_{ki}$$

where l is the number of significant feature vector pairs, f_k is a forecasting function for the kth inter-day feature vector, $\mathbf{u}_k$, such as the linear function in (5), and i_k^* is a set of historical lags which can be used to predict u_{ki}. Our current forecasting model (7) has $l = 1$ and $i_1^* = 1$.

In addition to call arrival data, another important part of call centre operational data is service data, which records how long each call lasts. We can foresee how SVD is applied to service data to identify the effects of time-of-day, day-of-week, service type, clusters of CSRs, and so on. Currently, a service time study is under way in combination with learning curve modelling of CSRs, where different summary statistics of the service data will be analysed, for example, service times' mean, median, standard deviation, and so on. We think that our SVD-based techniques are broadly applicable in analysing two-way structured data encountered in business and industry besides call centre data.

REFERENCES

1. Call Center Statistics. http://www.callcenternews.com/resources/statistics.shtml. 2002.
2. Utchitelle L. Answering '800' calls, extra income buy no security. *The New York Times* 2002, **March 27**: Section A, p. 1, Column 5.
3. Gans N, Koole G, Mandelbaum A. Telephone call centres: tutorial, review, and research prospects. *Manufacturing & Service Operations Management* 2003; **5**:79–141.
4. Brown LD, Gans N, Mandelbaum A, Sakov A, Shen H, Zeltyn S, Zhao L. Statistical analysis of a telephone call centre: a queueing-science perspective. *Journal of the American Statistical Association* 2005; **100**:36–50.
5. Mandelbaum A. Call centres, research bibliography with abstracts. *Technical Report*, 2003.
6. Avramidis AN, Deslauriers A, L'Ecuyer P. Modeling daily arrivals to a telephone call center. *Management Science* 2004; **50**(7):896–908.
7. Jongbloed G, Koole GM. Managing uncertainty in call centres using Poisson mixtures. *Applied Stochastic Models in Business and Industry* 2001; **17**:307–318.
8. Harville DA. *Matrix Algebra From a Statistician's Perspective*. Springer: New York, 1997.