A Method for Finding Similar Time Series and Forecasting with Calendar Constraints – a Commercial Bank Case Study

Krzysztof Kania
Knowledge Engineering Department
University of Economics
Katowice, Poland
krzysztof.kania@ue.katowice.pl

Jerzy Michnik
Operations Research Department
University of Economics
Katowice, Poland
jerzy.michnik@ue.katowice.pl

Abstract— **In many cases, especially in business activities we can observe constrains regarding an arrangement of the calendar. Sometimes they can be neglected, but sometimes they have a significant impact on the course of events. This article presents a simple method supporting the forecasts of this type of phenomena based on a concept of calendar similarity that may supplement traditional forecasting methods. Presented method has been used in the commercial bank to predict a volume of the documents to process.**

***Keywords** - forecasting; calendar; qualitative methods.*

I. Introduction

Business reality is always time-dependent and forecasting is one of frequently formulated tasks in successful management. Forecasting plays a special role when:

- resources cannot be stored (e.g., energy or work) or storage is very costly,
- shortages of resources may lead to distortions in the functioning of the organization or may lead to major losses,
- storing too large volume of resources creates a risk of wasting resources (a lack of jobs for workers, penalties for unused capacity),
- rapid change in the volume of resources is not possible or is costly.

In such cases, we need an accurate forecast early (sometimes even a few months earlier) and for every subperiod of the forecasted range (i.e., for every day of a whole month in advance). Some business activities and events exhibit variations depending on the specific day of the week or month. Therefore, they can be described as dependent on the calendar. Dependence of the predicted phenomena, events and processes with the calendar may be due to legal regulations or generally accepted norms and customs.

Examples of such phenomena include the different media consumption in the industry on some days (for example, before various holidays, days off, end of the year), the volume of traffic or the number of waiting customers. A similar problem also arises in planning of deliveries to large shops. Another example is volume and type of documents received in offices, banks and post offices during particular days connected with payments, deposits, withdrawals, transfers or the load on the servers for electronic services as well. The specificity of these phenomena is that the relatively stable long cycles (yearly, monthly, weekly) interfere with arrangement of weekdays, public holidays and additional days off.

Forecasting in such conditions requires consideration of factors that are disregarded in the analysis of phenomena that have uniform distribution in time or are insensitive to the arrangement of the calendar. Classical statistical methods do not include a quantitative prediction of the calendar. Methods for finding similarities in time series – very useful in many cases – do not account calendar directly and involved constraints as well. Hence, we need to use additional qualitative methods based on the large amount of data stored in data warehouses. The tasks of this class are in the scope of Business Intelligence systems and to implement them we use a variety of statistical tools and techniques such as neural networks and sequence analysis (see [1][2][3]). The paper presents a method to support prediction with taking calendar constraints under consideration with the example of the task of calculating the volume of processed documents.

The paper is organized as follows. In the next section (II) we define the problem raised in the bank and describe the goal of our research. In Section III, the whole procedure is outlined and presented in details using the example of one forecasted month. The paper ends with the short conclusions.

II. Problem definition

The method proposed has been developed for forecasting a number of people needed to process documents in one of the commercial banks and will be illustrated by the particular example. However, it seems that the method in question can be easily generalized and applied in other fields.

Bank branches receive traditional documents (transfers, fees, taxes, etc.) from their customers. The amount of work to be done by the bank staff depends on the volume of documents and the work structure (processing different types of documents requires a different amount of work). An important limitation is connected with the necessity of working within strict deadline (time of opening sessions of interbank payments). Our problem has to fulfill the following conditions:

- The forecast must be prepared at least one month earlier to plan holidays or prepare other work for employees and to minimize the number of people remaining in readiness to perform work.
- During every month there are two special days (10th and 15th), when the number of documents is the highest (dates for paying taxes and other fees).
- Strong influence of additional factors such as changes in commission fees, opening/closing of branches, changes in types of documents, and their structure, and introducing the new rules for an electronic exchange of information generates some need for cooperation with an experienced user in order to verify our predictions.
- It is necessary to take into account the distinction between weekdays and holidays.

The goal is to predict a number and type of documents to be processed in successive working days of the forecasted month as precisely as possible and, consequently, to determine a number of people needed for processing of the documents. For the procedure we assumed that:

- Total volume for the predicted period is known.
- Distributions of intensity of work from the historical data are known.
- Calendar arrangement causes large changes in the distribution of the phenomenon.

The total monthly volume of documents is predicted with the help of traditional statistical methods such as trend analysis, analysis of the relative and absolute deviations and analysis of the cycles of higher order (in this case - annual). In practice, these methods let us to predict very closely the total volume of documents for specific month.

III. Procedure for dividing the volume of work

The most difficult part of the whole research was to find the distribution of the total volume of documents for every particular day of the month. Fig. 1 shows the percentage distributions of the number of documents in the following days of four different months taken from historical data (gaps in the distributions relate to days off).

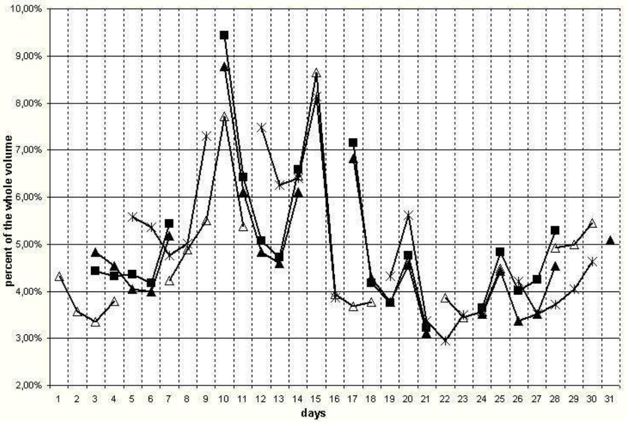


Figure 1. Sample month distributions of the intensity of the work

The monthly distributions of the phenomenon are dissimilar due to the different weekdays arrangement with respect to consecutive days of the month. In some months, the maxima fall to approximately 10th and 15th day, but in the other months distributions are different because the maxima fall on days off. In these cases the months maxima shift to or spread out on the preceding or following working day.

Due to these differences and gaps, a forecast based on the whole set of distributions leads to nowhere. For the same reason other methods of analyzing cyclic time-series like classical statistical analysis, Fourier analysis or wavelet analysis – effective in predicting of the continuous and uninterrupted time series – in this case turn out to be useless (see [4][5][6]). For that reason a new procedure has been proposed. Its outline is shown in Fig. 2.

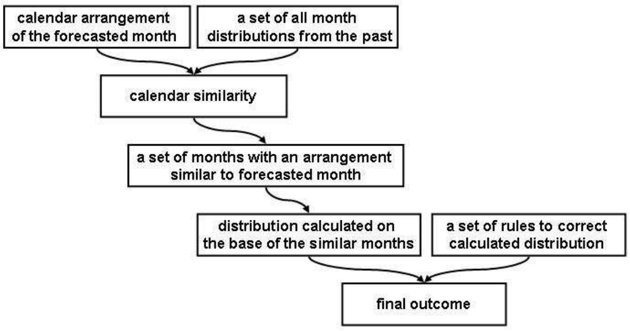

calendar arrangement of the forecasted month
a set of all month distributions from the past
calendar similarity
a set of months with an arrangement similar to forecasted month
distribution calculated on the base of the similar months
a set of rules to correct calculated distribution
final outcome

Figure 2. Outline of the procedure for finding distribution in a particular month

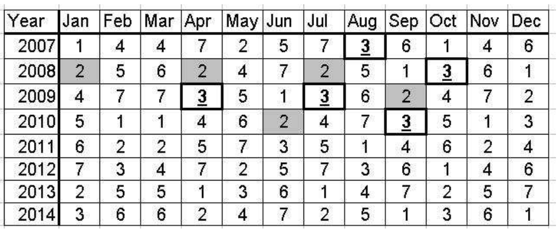
To distribute the volume of work over the working days of the particular month it is necessary:

- To find months with identical or similar calendar arrangement to the forecasted month.
- To incorporate Saturdays, Sundays and additional holidays.
- To improve the method with a set of rules correcting the initial distribution.

The procedure can be supported by Excel solution, so the whole knowledge needed to find similar months and to use additional rules was stored in 4 interrelated matrixes. It was decided to record information in the form of a matrix rather than in procedures or functions, as in a spreadsheet, matrices can be easily operated by users (e.g. introduction of a new day off) without additional tools.

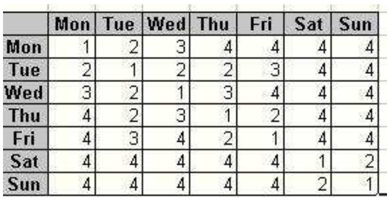
The first step of the procedure – determining the set of months that are similar in the calendar arrangement to the month of the forecast – is based on the content of the two first matrixes (Fig. 3 and Fig. 4). They are associated with a plain observation that the day of the week that starts the month determines the month arrangement until the 28th day. Months have different lengths but this is not significant as the analyzed phenomenon is not volatile at the very end of the month. That means, for example that February, although being 28 days long, may be in terms of a calendar arrangement similar to a longer month (June or July for example).

The second matrix (Fig. 4) contains ranks needed for finding in historical data months with identical or very similar calendar arrangement as in the predicted month. The rank equals to 1, in the matrix means that calendar arrangement of two months is identical (days off and special days 10th and 15th, in the same places). The rank equals to 2 means that calendar arrangement of two months is very similar but not identical, what causes changes in distribution and so on. Ranks are constant and was found through analyzing calendar, interviews with users supported by a graphical analysis of historical data and projections through the simulation.



Year	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct	Nov	Dec
2007	1	4	4	7	2	5	7	**3**	6	1	4	6
2008	2	5	6	2	4	7	2	5	1	**3**	6	1
2009	4	7	7	**3**	5	1	**3**	6	2	4	7	2
2010	5	1	1	4	6	2	4	7	**3**	5	1	3
2011	6	2	2	5	7	3	5	1	4	6	2	4
2012	7	3	4	7	2	5	7	3	6	1	4	6
2013	2	5	5	1	3	6	1	4	7	2	5	7
2014	3	6	6	2	4	7	2	5	1	3	6	1

Figure 3. A part of the matrix containing the number of weekday that starts a month (in Poland Monday is the 1st day of the week)



	Mon	Tue	Wed	Thu	Fri	Sat	Sun
Mon	1	2	3	4	4	4	4
Tue	2	1	2	2	3	4	4
Wed	3	2	1	3	4	4	4
Thu	4	2	3	1	2	4	4
Fri	4	3	4	2	1	4	4
Sat	4	4	4	4	4	1	2
Sun	4	4	4	4	4	2	1

Figure 4. The matrix of months' similarity rank

For example, in respect of the calendar, September 2010 (started at Wednesday) is the same as April and July 2009, October 2008, etc. (bolded and underlined in Fig. 3) as they starts with the same weekday and have rank equal to 1 in the matrix of months' similarity, and is very similar to June 2010, September 2009, January, April and June 2008 (shaded in Fig. 3) as these months started at Tuesday and the rank between Tuesday and Wednesday is equal of 2.

The result of this part of a procedure is a set of the months in which the distribution of intensity of work are actually comparable. In practice only months with rank 1 or 2 were used in forecasting because the distributions from the other months were too different. Since the calendar might be affected by some other factors (such as moving Easter or so called "long weekends"), the final decision on the choice of months is left to the analyst.

In fact, selecting only a few of many months decreases the basis of forecasting but on the other hand leaves only these months that have really similar distributions. Fig. 5 presents graphs for a selected month similar to September 2010. For the rest of the procedure we use an arithmetic mean from values for each day of the month (line with circles in Fig. 5).

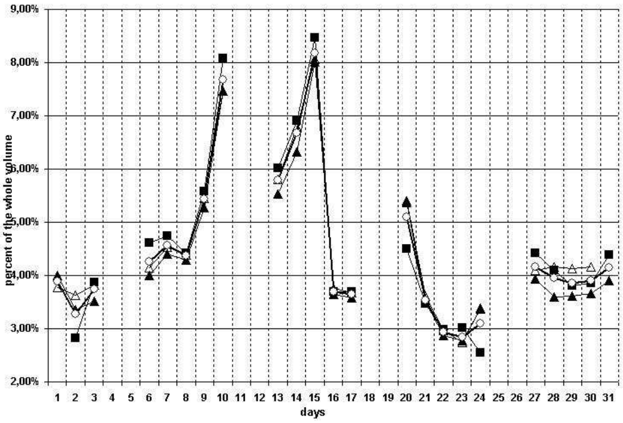
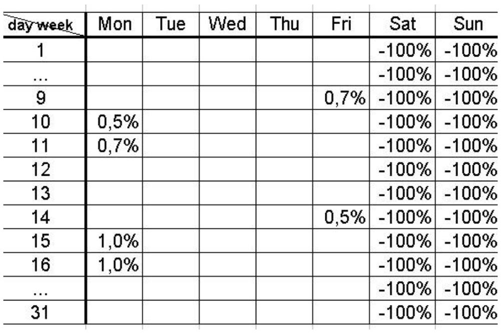


Figure 5. Distributions of selected months and a mean distribution (line with circles) of the intensity of the work for the forecasted month

As the distribution is obtained from different months (shorter and longer), the next step of the procedure is to verify and align the mean distribution obtained for the forecasted month. All verification rules have exactly two arguments in premise and exactly one value in the conclusion. This allowed for writing rules in the form of two-dimensional arrays. The arguments are: a day of the month and a day of the week (Fig. 6) or a month and a day of the month (Fig. 7). The element of the array is a percentage correction that should be made for the combination of arguments it belongs to. A user can use these correction rules to take into account additional factors that are concerned with particular days of week, month or year (Fig. 6 and Fig. 7). All the values in these matrices have been determined empirically and corrected on the basis of the experience of the bank staff, past observations and arrangement of the calendar for a current year.



day week	Mon	Tue	Wed	Thu	Fri	Sat	Sun
1						-100%	-100%
...						-100%	-100%
9					0,7%	-100%	-100%
10	0,5%					-100%	-100%
11	0,7%					-100%	-100%
12						-100%	-100%
13						-100%	-100%
14					0,5%	-100%	-100%
15	1,0%					-100%	-100%
16	1,0%					-100%	-100%
...						-100%	-100%
31						-100%	-100%

Figure 6. A part of the matrix of rules in the week-month relation

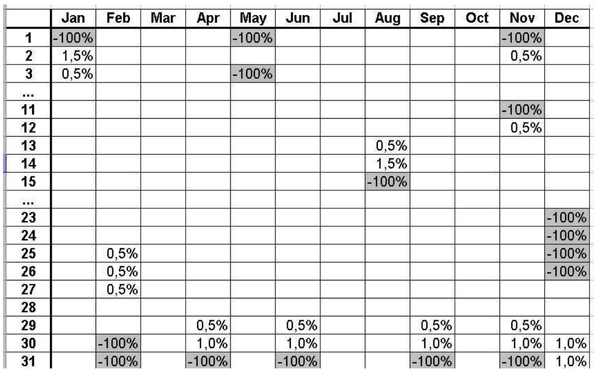
Information, contained in the matrix in Fig. 6, is presented to the user in the form of rules:

```
If    n-th day of the month falls on
                 particular weekday
then  change the value of that day by x%
```

For example, an element (9, Fri), of the matrix has a value 0.7%. This value corresponds to the rule:

```
If    the 9th day of month falls on
      Friday
then  increase the value of forecast on
      that day by 0.7%
```

This example reflects the knowledge that since the 10th (one of the special days) falls on Saturday, the volume of documents will be greater in the day before.



	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct	Nov	Dec
1	-100%				-100%						-100%	
2	1,5%										0,5%	
3	0,5%				-100%							
...												
11											-100%	
12											0,5%	
13								0,5%				
14								1,5%				
15								-100%				
...												
23												-100%
24												-100%
25		0,5%										-100%
26		0,5%										-100%
27		0,5%										
28												
29				0,5%		0,5%			0,5%		0,5%	
30		-100%		1,0%		1,0%			1,0%		1,0%	1,0%
31		-100%		-100%		-100%			-100%		-100%	1,0%

Figure 7. A part of the matrix of rules in the year-month relation

The last matrix (Fig. 7) contains values for rules correcting each day of the year especially due to holidays or the different length of the months. At the end of shorter months the volume of documents grows in relation to the mean distribution. Similarly, higher intensity of work is observed right before or after holidays (for example, in Poland August 15th or November 1st). These values are also presented to the user as suggestions and they are as follows:

```
On September 29, the system proposes to
      increase the value by 0.5%.
On September 30, the system proposes to
      increase the value by 1.0%.
On September 31, the system proposes to
      decrease the value by 100.0%.
```

These three particular rules show that the volume of work in September (shorter month) shifts to two previous days.

A user may accept or reject proposals adjusting the distribution to get the final form. Sometimes using the rules may result in the situation that the sum of intensities of work moves away from 100%. In that case an analyst can also manually make changes to the proposed schedule increasing or decreasing all the values throughout the forecast period.

Fig. 8 shows the final result of this procedure – the distribution of the intensity of work during the entire month. Comparing it with the mean distribution we can see that it is slightly different due to correcting rules (circled parts of the distribution).

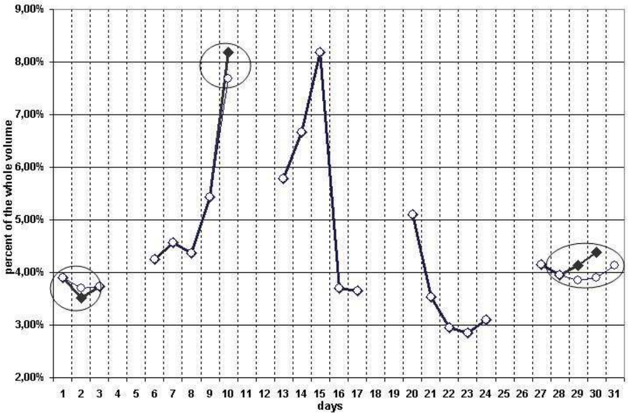


Figure 8. Corrected distribution for September 2010

At the last phase of the procedure, we calculate a number of documents for each day of the forecast by dividing the total monthly volume of documents according to obtained distribution. And finally as an average workers' loading is known we can calculate a number of people needed to process the documents for every day.

IV. CONCLUSIONS

The presented method has become a module of a larger system that was implemented in the commercial bank. The procedure described above, has replaced previously used forecasting method based on a simple analogy and improved its results. Moreover, the whole procedure was improved with mechanism for storing data and forecasts in the database. This allowed to connect forecasts with the scenario method and to easily conduct the what-if analysis. Currently, due to introduction of electronic banking, a number of traditional documents processed has declined significantly but it seems that the proposed method could be used for a prediction of other phenomena whose course depends on the specific arrangement of the calendar.

REFERENCES

[1] Adamo J.M.: Data Mining for Association Rules and Sequential Patterns, Springer-Verlag, New York, 2001.

[2] Han J. and Kamber M.: Data Mining Concepts and Techniques, Academic Press, 2001.

[3] Kovalerchuk B. and Vitayaev E.: Data Mining in Finance, Kluwer Academic Publishers, 2000.

[4] Kania K.: "A New Measure and Symbolic Method that Supports Finding Similarity in Time Series", in: Business Information Systems, W. Abramowicz, G. Klein (eds.), Colorado Springs, USA, 2003, pp. 124-131.

[5] Caraca-Valente J. and Lopez-Chavarrias I.: "Discovering Similar Patterns in Time Series", 6th ACM SIGKDD Int'l Conf. on Knowledge Discovery and Data mining, Boston, 2000, pp. 497-505.

[6] Keogh E. and Pazzani M.: "A simple dimensionality reduction technique for fast similarity search in large time series databases", 4th Pacific-Asia Conf. On Knowledge Discovery and Data Mining, Kyoto, 2000, pp. 122-133.