Data mining in publishing: a nice feature or a necessity?

T. M. Fernández-Steeger[1], F. Zander[1], S. Callsen[1], S. Steinberg[1] & N. Brauns[2]
[1]*VKG Verlagvertriebs KG – Bauer Verlagsgruppe, Hamburg, Germany*
[2]*University of Applied Sciences Wedel, Wedel, Germany*

Abstract

The two distribution channels in the publishing business – retail and subscription – compete against but at the same time complement each other. The following three examples show the importance and potentials of data mining in this stress field. The first example shows a model using Markov chains that will explain the coherence of these units. The other examples show how to estimate the medium lifetime of subscriptions and subsequently the maximum IPO. The last example finally explains how to create test groups for market research by employing supervised clustering and points out the advantages of using this method.
Keywords: data mining, publishing, marketing, subscription, Markov chains, maximum IPO, Tornquist function, supply test, supervised clustering.

1 Introduction

Bauer Verlagsgruppe (Bauer Publishing Group), founded in 1875, is one of the largest journal publishers in Europe. The company publishes 120 journals in 30 countries (Europe and overseas). The total sales in 2002 reached Euro 1,71 billion, with a percentage of 63% for Germany. The VKG Verlagvertriebs KG is the sales and marketing company of the Bauer group for German-speaking journals and provides about 50% of the group sales. The data mining department supports the VKG main units, the retail and the subscription unit.

As today's journal market in Germany is a nearly mature and strongly regulated market, it becomes more and more difficult for the main market participants to generate further growth without negative effects on their financial overall result. This is even more relevant for the market leaders. For that reason,



Data Mining V, A. Zanasi, N. F. F. Ebecken & C. A. Brebbia (Editors)
© 2004 WIT Press, www.witpress.com, ISBN 1-85312-729-9

a clear understanding of the market and its interactions is pre requisite. Data mining techniques are an excellent support to act more efficiently in this difficult environment. The following examples show in which way data mining may cause better understanding and controlling of company and market processes.

2 Interaction between retail and subscription

A controversial discussion in the publishing business is provoked by the interaction of retail and subscription sales. Retail customers may become subscription customers and vice versa. But is it worth investing money to convert a retail customer into a subscription customer? Furthermore, which maximum investment per order (IPO) can be accepted for that conversion? Usually in Germany retail and subscription business are independent departments or even separated companies. For this reason they often compete with each other and dispute whether subscription create additional profit in sales or not.

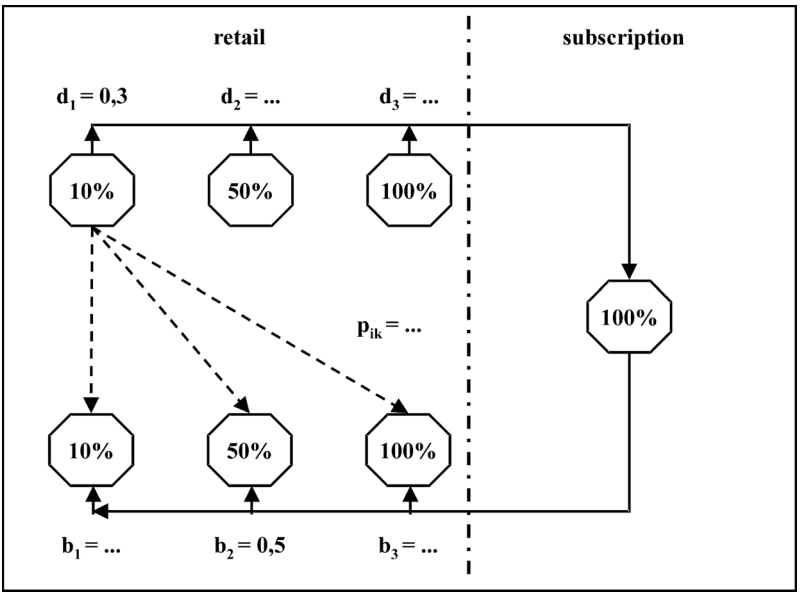


Figure 1: Model to predict retail and subscribtion sales by time.

2.1 Model retail / subscription interaction

Normally the profit of a subscription is determined independently without taking into account the retail interaction. The retail department criticises this approach, as in their opinion subscriptions create the same overall sales volume using different distribution channels. From their point of view the additional investments for subscription business are inefficient.

Data Mining V, A. Zanasi, N. F. F. Ebecken & C. A. Brebbia (Editors)
© 2004 WIT Press, www.witpress.com, ISBN 1-85312-729-9

In fact both views are wrong, as they neglect the interaction and migrations between this groups. Often subscription customers have been retail customers before. Furthermore not all subscribers quitting their subscription stop buying those journal. Moreover retail customers do not purchase every issue. Finally the buying pattern of customers changes occasionally.

Considering these assumptions the model in figure 1 was developed. A group of potential subscription customers will be divided in a retail group and a subscription group. Modeling the retail yield, Markov chains (e.g. Carter [1]) were used (left part of the model). The retail group is subdivided into groups corresponding to their buying frequency (octagons) with a given start distribution expressed by the vector D. Taking care of the changing behaviour of the customer, interchanging probabilities (p_{ik}) were taken into account by the interchange vector P. The expected discounted yield S_{ret} for a retail customer can be obtained from the following equitation (r = rate of interest, N = planning interval in years, Q = buying frequency):

$$S_{ret} = D * \sum_{t=1}^{N} \frac{P^t}{(1+r)^t} * Q \tag{1}$$

For members of the subscriber group the changes from and to retail have to be taken in consideration. For the subscription sales S_{sub_sub} (eqn. 2) the start vector D is not needed, as subscribers buy all issues. After quitting the subscription, the vector B describes the new distribution of the purchase frequency. This is necessary to determine the retail sales from former subscribers S_{sub_ret} (eqn. 3). The discounted yield for members (eqn. 4) of the subscriber group can be described by the following equations, in which A is the annual return and k the mortality rate.

$$S_{sub_sub} = \frac{A}{r}(1 - (\sum_{t=1}^{N-1} \frac{k_t}{(1+r)^t}) + \frac{1}{(1+r)^N}((\sum_{t=1}^{N-1} k_t) - 1)) \tag{2}$$

$$S_{sub_ret} = B * [\sum_{t=1}^{N-1} (\frac{k_t}{(1+r)^t} \sum_{\tau=1}^{N-t} \frac{P^\tau}{(1+r)^\tau})] * Q \tag{3}$$

$$S_{sub} = S_{sub_sub} + S_{sub_ret} \tag{4}$$

The medium lifetime (MLT) of the subscription can be derived from the following equitation using the mortality rate k.

$$H \approx \sum_{t=1}^{N-1} k_t t \tag{5}$$

2.2 Simulations and reality

The modeling does not show only a close relationship between MLT and the yield of the subscription. The overall return depends exceedingly on the distribution vectors B/D and the interchange vector P. In table 1 an example for

Data Mining V, A. Zanasi, N. F. F. Ebecken & C. A. Brebbia (Editors)
© 2004 WIT Press, www.witpress.com, ISBN 1-85312-729-9

the comparison between retail and subscription sales can bee seen. The distribution vectors B and D where approximated using real data.

This example shows that the subscription business creates additional sales. The sales plus is nearly equivalent to the subscription sales (-5%). The error of about 5% is negligible as the influences of other factors as e.g. prices, taxes and economic cycles are even higher.

Table 1: Sold issues per customer on the distribution channels in a given period of time.

Observation period in years	Retail	Subscription		Additional sales
	retail sales	subscription sales	retail sales	
10	74	141	67	134
15	92	148	84	140
20	103	151	95	143
25	111	152	103	144
30	115	153	107	145

Besides this simulation extensive market research and surveys were made to prove this results for customers of two different journals, a food magazine and a woman journal. The results where comparable to the simulations described before. In both cases the additional sales were again nearly equivalent to the subscription sales. In case of the food magazine the error was about 6%. In case of the woman weekly the error was 3%. This means that the subscription business makes sense and the separate determination of subscription and retail business is acceptable.

3 Lifetime value and IPO

As seen above it is an additional benefit for the company to convince their customers to take out a subscription. Thus investments are necessary to convince customers to become subscribers. Therefore detailed information about the expected maximum IPO (investment per order) are necessary. To calculate this IPO, the MLT of the subscription is pre-requisite. This is even more important for new or younger products where only rare database are available.

3.1 Methods for lifetime estimations

The usual way to evaluate the subscriber's lifetime is the same as used for the calculation of the lifetime of humans. The mortality rate is calculated for every journal and any subscription period. Finally a specific loss function for every

Data Mining V, A. Zanasi, N. F. F. Ebecken & C. A. Brebbia (Editors)
© 2004 WIT Press, www.witpress.com, ISBN 1-85312-729-9

journal is obtained (fig. 2). The surface below this function is the MLT which can be derived from the following equation (s. Hartunge [2]).

$$F_X(t) = 1 - e^{-\int_0^t r(\xi)\,d\xi} \tag{6}$$

Actually there are subscription periods longer than 30 years, but from an economical point of view those periods are negligible. For this reason the maximum subscription period is limited to 30 years.

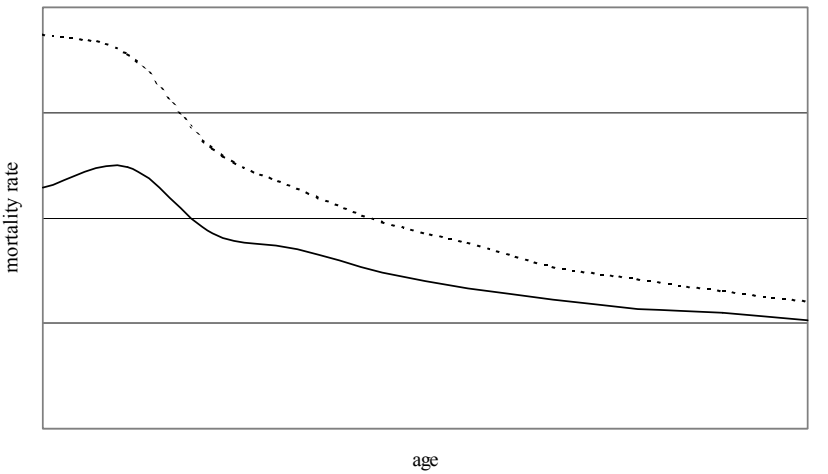


Figure 2: Potential loss function for two journals.

Table 2: Errors between standard and tornqvist lifetime.

Journal	standard	Tornqvist	error
one	4,75	4,63	2,50%
two	5,73	5,62	2,00%
three	6,3	6,36	0,90%
Four	6,61	6,7	1,50%

This easily works for approved and successful journals with a long market presence. For new or younger journals this method can not be applied, as there are insufficient observations. Therefore two methods where developed to solve the missing value dilemma. For younger journals with sufficient observations between 3 to 5 years the Tornqvist function (eqn. 7) is used to extrapolate the mortality rate.

$$H_t = \frac{a(t-1)}{(t-2) + a/b} + c \tag{7}$$

© 2004 WIT Press, www.witpress.com, ISBN 1-85312-729-9

The comparison between the standard method and the Tornqvist function shows a good accordance(s. fig. 3) and an acceptable error rate (tab. 2).

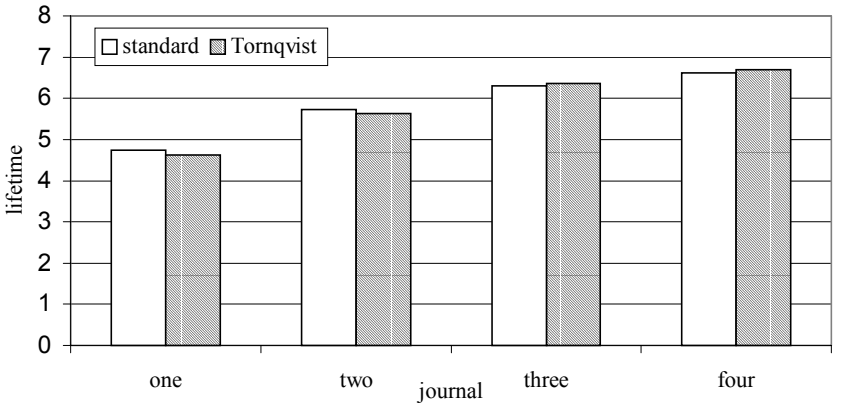


Figure 3: Comparison of standard lifetime and tornqvist lifetime.

For new journals where insufficient data is available, the Tornqvist method cannot be applied. Therefore another method is used. In that case data of comparable journals is used to build up a database.

3.2 IPO and discounted cash flow method (DCF)

To calculate the maximum IPO, the discounted cash flow method has been modified. Usually the equation below is used to calculate the DCF (s. Reichmann [4]).

$$\mathrm{DCF} = -a_0 + \sum_{t=0}^{n} \frac{e_t - a_t}{(1+i)^t} \tag{8}$$

Due to the fact that there is no prior information of the real lifetime of every subscription, the economic life n (eqn. 8) of the investment was replaced by the MLT. This leads to a non realistic description of the cash flow. Therefore a modified DCF equitation is used to reproduce a realistic cash flow. The figure below (fig. 4) shows the comparison between maximum IPOs for different rates of interest, using real discounted cash flow's and the modified cash flow model. The maximum error rate is within range of usual tolerance limits.

This predictive model is crucial for investment strategies and decisions in the publishing business. It also builds the basis for marketing controlling, such as ex-post analysis of marketing campaigns.

Data Mining V, A. Zanasi, N. F. F. Ebecken & C. A. Brebbia (Editors)
© 2004 WIT Press, www.witpress.com, ISBN 1-85312-729-9

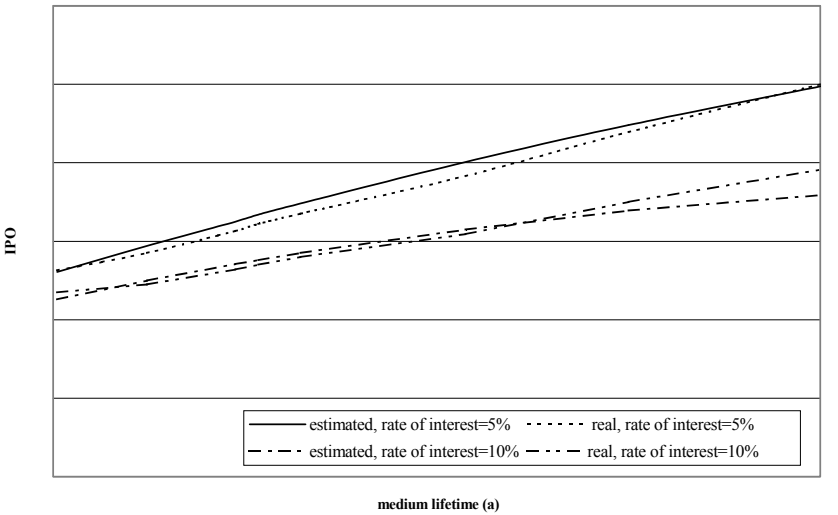


Figure 4: Comparison of IPO's calculated with real and modified cash flow.

4 Supply policy/regulation

4.1 Theoretical considerations

Another field of investigation concerning retail trade analyses is the field of optimal supply of products. The optimal supply has always be regarded in relation to the actual demand.

Consider the relation between supply and demand depicted below. This well-known context can be found in any basic economic textbook and has two important characteristics: Obviously, the higher the supply, the more you sell. Secondly, with increasing supply the additional sales diminish.

The optimisation problem is to choose the point on the curve where marginal supply and marginal sales arise, so that marginal cost (MC) equal marginal revenues (MR). Microeconomic theory shows that this is the first order condition for maximum profits. If we denote P as profits, R as revenues and C as costs, all as functions of given supply x the following can be written:

$$P(x) = R(x) - C(x) \tag{9}$$

The first order condition for maximum profits is the derivative of the profit function with respect to x:

$$MP(x) = MR(x) - MC'(x) = 0 \tag{10}$$

$$MR(x) = MC(x) \tag{11}$$

© 2004 WIT Press, www.witpress.com, ISBN 1-85312-729-9

4.2 Testing

Now there are two questions of interest: where is the observation point on the curve (fig. 5) and what marginal cost and marginal revenues exist in this situation? In order to find answers to this question every company varies their supply to some extent. The decision rule sounds very easy: if we increase our supply B (s. fig. 5) and additional costs are higher than additional revenues, than go back to your initial situation A. Correspondingly, we are able to reduce output C and compare sinking costs to decreasing revenues and apply the same decision rule.

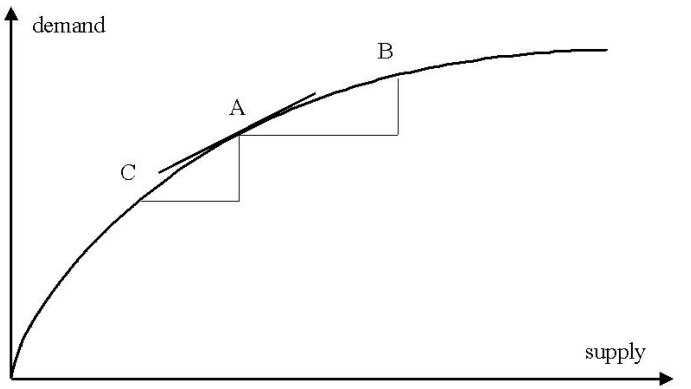


Figure 5: Theoretical supply curve.

In reality, things are not so easy as they seem. First of all, the situation chosen in the past should be, in general, near to the optimum, so that supply variations for testing different situations are carried out in small amounts, say +/- 2 per cent. Secondly, the supply and demand function does not really look like the ideal curve above (fig. 5), but rather as the curve depicted below (fig. 6). This is due to the sum of many random effects, for instance economic fluctuations, changing market conditions, actions of competitors, weather conditions and so on.

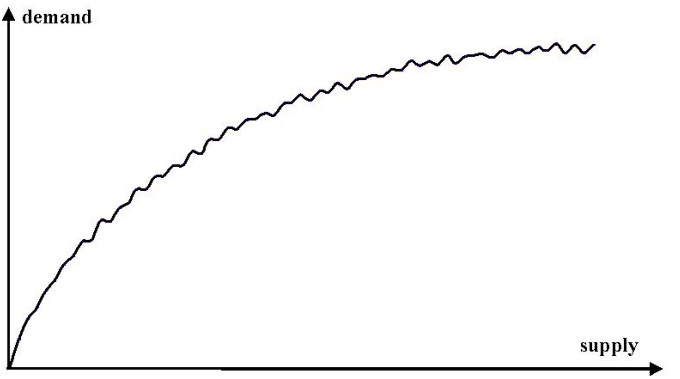


Figure 6: Real supply curve with random effects.

Data Mining V, A. Zanasi, N. F. F. Ebecken & C. A. Brebbia (Editors)
© 2004 WIT Press, www.witpress.com, ISBN 1-85312-729-9

The consequently whenever testing new situations the additional revenues caused by output variations are heavily distorted and the measurement of additional effects is much more difficult. In the worst case, the (real) additional effects are smaller than the random effects.

4.3 Supervised clustering

The market for publishing papers in Germany is splitted into roughly one hundred regions. Tests of output variations are performed only in selected parts of Germany, assuming ten per cent of all regions.

The data mining approach in the Bauer publishing group could be described as the following: take ten out of hundred regions, so that the sum of all random effects balance out and become minimal (Pohl & Steinberg [5). For example: one region has two peaks, the second region has two dips at the same time, so the overall sales of these two regions remain constant in time(s. fig. 7).

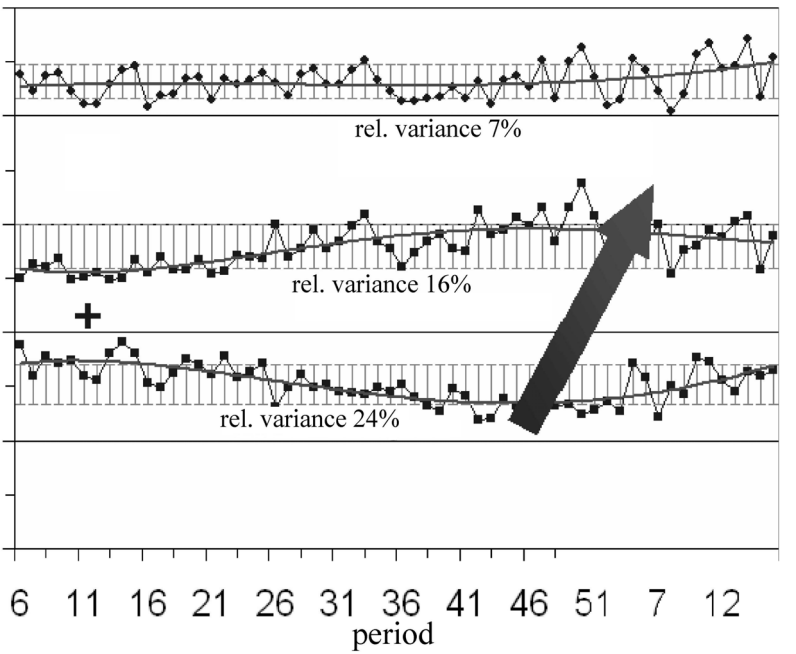


Figure 7: Balancing regional sales (supervised clustering).

The table below (tab. 3) shows the mean deviation of sales for five groups of regions which were grouped by supervised clustering. In contrast 42 regions were not grouped at all and show significantly higher mean deviations. The clustering procedure therefore assures minimal oscillated time series of sales and ameliorate measurements of test scenarios.

Data Mining V, A. Zanasi, N. F. F. Ebecken & C. A. Brebbia (Editors)
© 2004 WIT Press, www.witpress.com, ISBN 1-85312-729-9

Table 3: Medium deviations of sales for grouped and non grouped regions.

Journal X	Number of regions	Medium deviation of sales			
		mean standard deviation in % (1 period)	mean standard deviation in % (6 periods)	absolute deviations (1 period)	absolute deviations (6 periods)
Cluster 1	9	0,31%	0,18%	373	219
Cluster 2	10	0,29%	0,18%	344	211
Cluster 3	10	0,42%	0,33%	505	401
Cluster 4	13	0,58%	0,50%	695	602
Cluster 5	11	0,28%	0,17%	340	202
Not clustered regions	42	0,97%	0,76%	392	352
All	**95**	**0,37%**	**0,28%**	**432**	**347**

5 Discussion

Coming back to the question of the heading "Data mining in publishing – a nice feature or a necessity" the definite answer will be "Companies can exist without data mining units". Thus the examples above show they can live better or more successful with data mining. More detailed information about companies processes and their interaction can be uncovered as the model for the retail subscription interaction shows. But also practical solutions for problems in the operative business can be provided for the operating departments as the examples for the maximum IPO approximation and the supervised clustering shows.

References

[1] Carter, E. F., *Random Walks, Markov Chains and the Monte Carlo Method*, Taygeta Scientific Inc, http://www.taygeta.com/rwalks/node7.html

[2] Hartunge, J., *Statistik – Lehr- und Handbuch der angewandten Statistik*, Oldenburg: Munich and Vienna, pp. 218-219, 1999.

[3] Tornqvist, L., The Bank of Finland's Consumption Price Index , *Bank of Finland Monthly Bulletin*, **10**, pp. 1-8, 1936.

[4] Reichmann, T., *Controlling mit Kennzahlen und Managementberichten*, Vahlenverlag: Munich, pp. 237-240, 1997.

[5] Pohl, S. & Steinberg, S., Heuristische Verfahren zur Aggregation addierbarer Zeitreihen. *Data Mining und Statistik in Hoschschule und Wirtschaft – Proceedings der 7. KSFE, Potsdam*, eds. Redlich, H. & Becker, C., Shaker Verlag: Aachen, pp. 157-165 2003.

[6] Hastie, T., Tibshirani, R. & Friedman, J., *The Elements of Statistical Learning – Data Mining, Inference and Prediction*, Springer: New York and Berlin, pp. 453-472, 2001.

© 2004 WIT Press, www.witpress.com, ISBN 1-85312-729-9