APPENDIX (2)

MARKET BASKET ANALYSIS

History

Market Basket Analysis (MBA) is the practice of finding out what products customers typically buy together. This knowledge is valuable:

- ⇑ In retail outlets, it can aid store layouts. "Impulse buy" products can be placed next to "regular buy" products with which they are associated. "Regular buy" products can be place some distance from associated "regular buy" products, encouraging the customer to tour the entire store.
- ⇑ Cross-promotional opportunities can be identified. In retail environments, special offers can be made to encourage product purchasers to buy associated products. In other industries, MBA can be used to identify which customers would be the best focus of marketing campaigns.
- ⇑ "Cause and effect" relationships can be identified, making it possible to predict what a customer is likely to buy next.

Until recently, MBA was only justified in the largest of companies due to the effort of collecting the data. This has changed dramatically for two reasons:

- ⇑ Barcode scan data automatically records the required information and can exploit it in real time at the cash register in the form of targeted promotions printed on the purchaser's receipt.
- ⇑ Web sites automatically record purchase data and can exploit it by immediately presenting the surfing customer with targeted product recommendations. A classic example of this is *Amazon.com*'s web-site. As the customer looks at the description of a product, a paragraph is added which says, "People who bought this product also bought..."

MBA is achieved by analyzing sales transaction data. Additional factors, such as region, channel, day of week (or time of year) and demographic factors about the customer can enrich the analysis.

Market Basket Analysis In Practice

MBA is usually implemented in-house rather than with off-the-shelf software. Since this is the case, this discussion will include a simple example. Any reader wishing to implement MBA can test their system with these numbers.

The sample data considers six products bought in twenty transactions:

	Quantity Bought					
Transn.	**Juice**	**Tea**	**Coffee**	**Milk**	**Sugar**	**Pop**
1	**0**	**0**	**0**	**0**	**0**	**0**
2	**0**	**2**	**2**	**4**	**3**	**0**
3	**1**	**0**	**0**	**0**	**0**	**0**
4	**0**	**1**	**0**	**0**	**0**	**0**
5	**1**	**2**	**1**	**1**	**0**	**0**
6	**0**	**2**	**1**	**3**	**2**	**0**
7	**0**	**0**	**0**	**0**	**0**	**6**
8	**0**	**0**	**0**	**0**	**0**	**0**
9	**4**	**0**	**0**	**0**	**0**	**0**
10	**0**	**0**	**1**	**1**	**0**	**0**
11	**0**	**0**	**0**	**0**	**0**	**6**
12	**0**	**0**	**1**	**1**	**0**	**0**
13	**0**	**0**	**0**	**0**	**0**	**5**
14	**0**	**0**	**0**	**0**	**0**	**0**
15	**1**	**2**	**0**	**2**	**0**	**0**
16	**0**	**1**	**1**	**1**	**2**	**1**
17	**1**	**0**	**1**	**0**	**0**	**0**
18	**2**	**0**	**0**	**0**	**0**	**0**
19	**0**	**0**	**0**	**0**	**0**	**2**
20	**3**	**0**	**0**	**0**	**0**	**3**

MBA looks at transactions to see which products get bought together. To do this, the transactions are analyzed to calculate:

N, the total number of orders.
n_i, the number of orders in which product i is bought.
x_{ij}, the number of orders in which both products i and j are bought.

In the case of this example, N is 20. n_i and x_{ij} are therefore:

	Juice	**Tea**	**Coffee**	**Milk**	**Sugar**	**Pop**
Ni	**7**	**6**	**7**	**7**	**3**	**6**

Xij	**Juice**	**Tea**	**Coffee**	**Milk**	**Sugar**	**Pop**
Juice	**7**	**2**	**2**	**2**	**0**	**1**
Tea	**2**	**6**	**4**	**5**	**3**	**1**
Coffee	**2**	**4**	**7**	**6**	**3**	**1**
Milk	**2**	**5**	**6**	**7**	**3**	**1**
Sugar	**0**	**3**	**3**	**3**	**3**	**1**
Pop	**1**	**1**	**1**	**1**	**1**	**6**

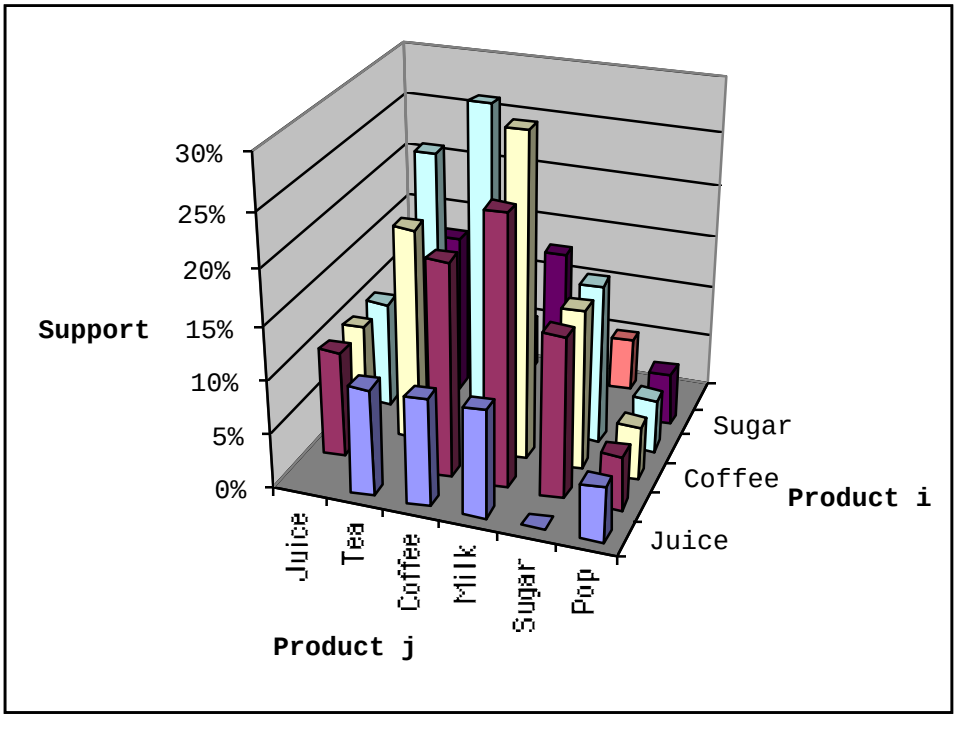
From this data, four measures of product association can be calculated: support, confidence, improvement and value.

Support

$$S_{ij} = \frac{x_{ij}}{N} \infty 100\%$$

Support	Juice	Tea	Coffee	Milk	Sugar	Pop
Juice		10%	10%	10%	0%	5%
Tea	10%		20%	25%	15%	5%
Coffee	10%	20%		30%	15%	5%
Milk	10%	25%	30%		15%	5%
Sugar	0%	15%	15%	15%		5%
Pop	5%	5%	5%	5%	5%	

Note that the leading diagonal values, which would always be 100%, are meaningless and have been omitted. Also note that the table is symmetric across the diagonal. For reporting purposes, this information is often best left as a table, but for comparison with the other measures, let us show it as a chart:



Confidence

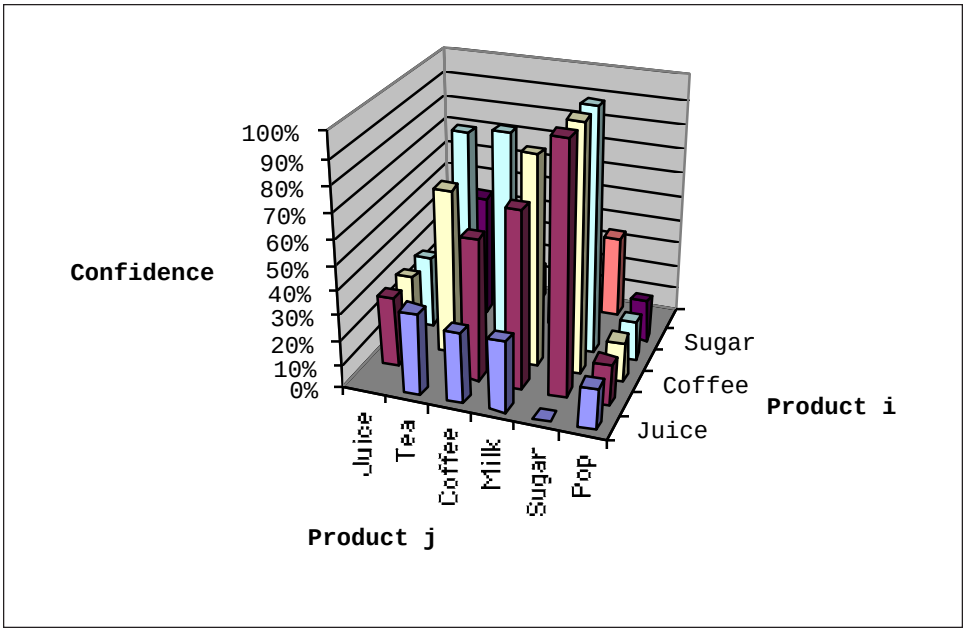
Confidence $C_{i\varnothing j}$ measures the percentage of buyers of product *i* who also buy product *j* and ıs calculated as:

$$C_{i\varnothing j} = \frac{x_{ij}}{n_i} \infty 100\,\%$$

For the data in the example, this gives:

		Product j					
		Juice	**Tea**	**Coffee**	**Milk**	**Sugar**	**Pop**
Product i	**Juice**		**33%**	**29%**	**29%**	**0%**	**17%**
	Tea	**29%**		**57%**	**71%**	**100%**	**17%**
	Coffee	**29%**	**67%**		**86%**	**100%**	**17%**
	Milk	**29%**	**83%**	**86%**		**100%**	**17%**
	Sugar	**0%**	**50%**	**43%**	**43%**		**17%**
	Pop	**14%**	**17%**	**14%**	**14%**	**33%**	

Note that the table is not symmetric across the diagonal: the average Sugar buyer is twice as likely to buy Pop as vice versa. For web-site cross-selling, the confidence measure can be used to inform the site visitor of other products they might be interested in. For purposes of comparison, the confidence data is graphically depicted as follows:



Improvement

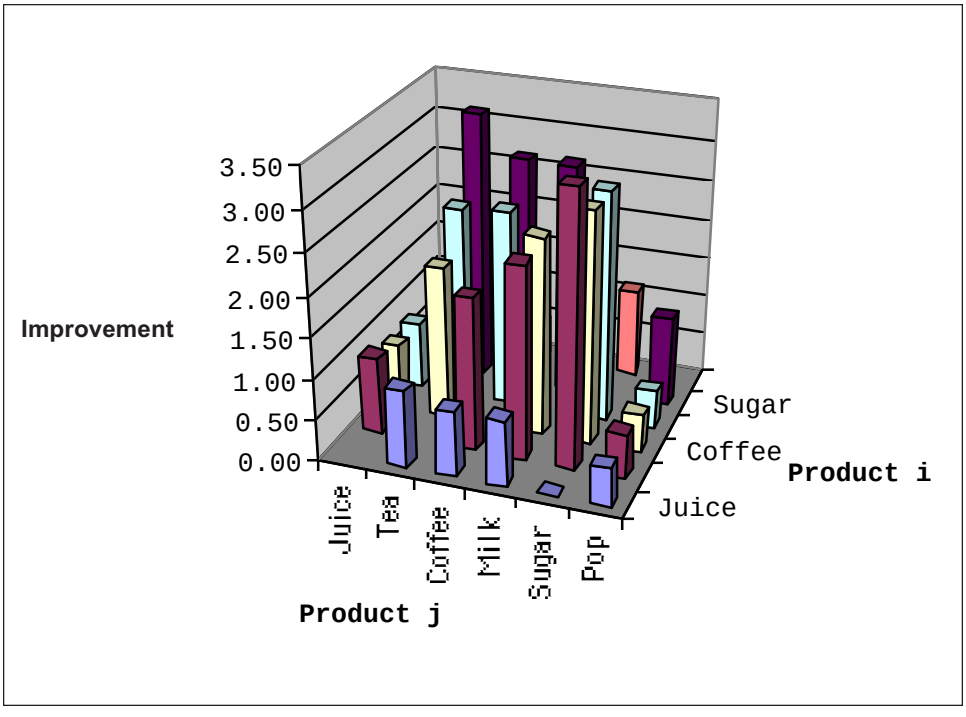
Improvement I_{ij} measures how much more likely product *i* buyers are to buy product *j* than customers in general:

$$I_{ij} = \frac{N \in x_{ij}}{n_i \in n_j}$$

For the data in the example, this gives:

Improvement	Juice	Tea	Coffee	Milk	Sugar	Pop
Juice	0.95	0.82	0.82	0.00	0.48	
Tea	0.95		1.90	2.38	3.33	0.56
Coffee	0.82	1.90		2.45	2.86	0.48
Milk	0.82	2.38	2.45		2.86	0.48
Sugar	0.00	3.33	2.86	2.86		1.11
Pop	0.48	0.56	0.48	0.48	1.11	

Surprisingly, this table is symmetric across the diagonal. This measure is commonly used to identify cross-selling promotional opportunities. It identifies which promotion is most likely to appear to the customer. For example, a sugar buyer is nearly three times as likely to want milk than the average customer. For purposes of comparison, the improvement data is graphically depicted as follows:



Value

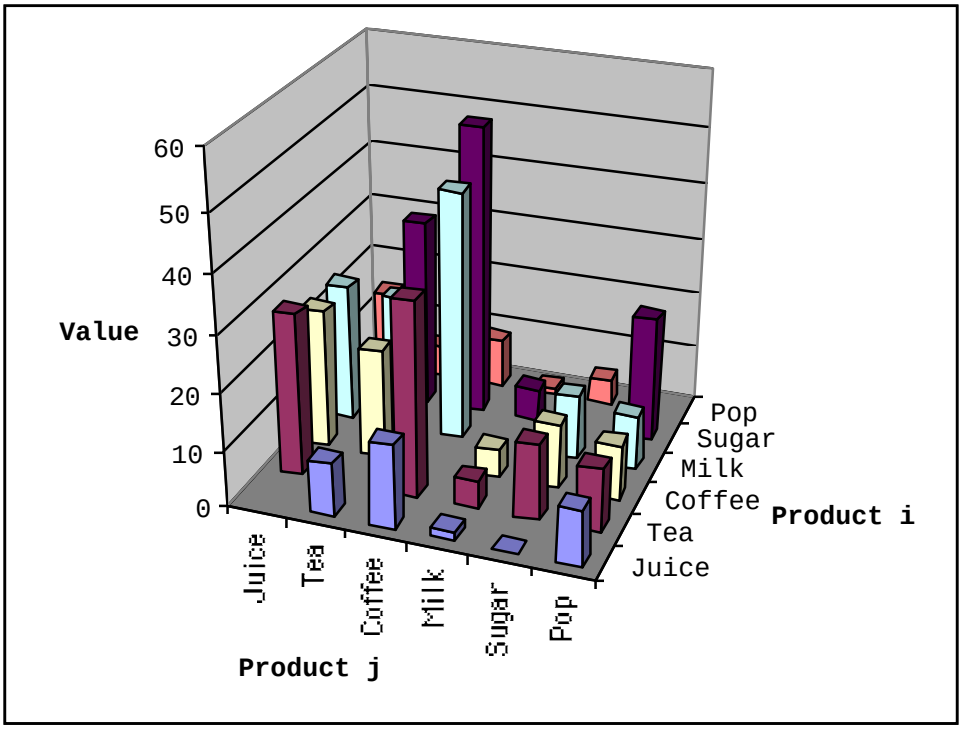
Value is a similar measure to improvement except that it measures the cross-selling potential in dollars profit rather than simple unit sales:

$$V_{i \varnothing j} = m_j \times I_{ij}$$

For the data in the example, this gives:

Value	Juice	Tea	Coffee	Milk	Sugar	Pop
Juice		10	15	2	0	10
Tea	29		34	5	13	11
Coffee	24	19		5	11	10
Milk	24	24	44		11	10
Sugar	0	33	51	6		22
Pop	14	6	9	1	4	

Because the margins of the promoted products have been incorporated, the table is no longer symmetric across the diagonal. This measure identifies the most profitable cross-selling opportunities. For example, promoting sugar to coffee buyers has the highest value payoff. This measure can be used to generate targeted discount coupons. Incorporating the margin information directs us clearly to the most profitable opportunities, as the chart shows:



Note the difference between the improvement and value measures: sugar and milk are low margin, and so the value measure does not attempt to cross sell them even though an attempt to cross-sell would probably be successful.

BEYOND BASIC MBA

There are many ways to extend this analysis, depending on the products being sold:

- Look for links between three or more products. This is useful for identifying product bundles, such as popular mixes of ingredients in a sandwich. In practice, this approach increases calculation times exponentially and has limited payoff.
- Take account of the quantity of product ordered in each purchase.
- Segment the analysis by region, channel, day of week (or time of year) or demographic factors about the customer, etc.
- If customer identity is recorded in the transaction detail, the analysis can be used to link products bought in this order with products bought in subsequent transactions.

Market Basket Analysis is discussed further in Berry (1997).