**V International Symposium Engineering Management and Competitiveness 2015 (EMC 2015)**

June 19-20, 2015, Zrenjanin, Serbia

Similarity Searching model with Excel

Zoltán Varga

Szent István University, Gödöllő, Hungary

zvarga87@gmail.com

**Abstract**

The here presented hybrid model is based on the available form of similarity searching in Excel. The baselines of the new method are to create sequences of mutually exclusive and non-overlapping categories, which generate different category-sequences, and by the logic of VLOOKUP the next of most similar pattern can be identified as the first element of a sufficiently accurate forecast. The estimations were evaluated by hit ratio and directional stability. In particular, testing the model on stock market time series, the results, achieved on the disaggregated minutes of open values of NASDAQ, with 290 minutes used to forecast the next 194 achieved 73.2% of directional stability, with 7 categories took into account, 55 cases were direct hit (28,35%), 86 cases were +/-1 category difference (44,33%), and 36 cases +/-2 category differences (18,46%). The step between the categories was 0,0000982 point. Testing it on sales time series of an industrial company’s product, the results achieved on the disaggregated weeks, with 100 weeks used to forecast the next 50 achieved 74% of directional stability, with 7 categories took into account, 24 cases were direct hit (48%), 17 cases were +/-1 category difference (34%), and 4 cases +/-2 category differences (8%). The step between the categories was 0,8276 point.

**Key words**: similarity, VLOOKUP, stock market, demand, forecast

**Introduction**

Following the Symposium of EMC 2014 new datasets were tested on a previous forecast model developed to solve sales and inventory problems in MS Excel. The basic illustration of the classical model is shown below on Figure 1.

 0. Historical sales data

1. Periodical simple average

2. Smooth step 1 by moving average

3. Smooth step 2 by exponential cleaning

4. Estimation

5. Forecasted values

6. Interpretation to the actual data

1. Figure: Classical forecasting model illustration

The estimation was carried out by the FORECAST function, which was not designed to deal with such datasets. Originally it had to have a dependent and independent range of data. In this particular case, it uses the values given after the exponential cleaning as an independent, and the values from the periodical simple average as dependent ones. This function also uses an X value, as a starting point, so it was determined as the last known data.

This model was tested on stock market datasets of Crude Futures weeks on the four main categories (Open, High, Low, Close). It was aiming to answer how universal this model is. The accuracy results of forecasting Crude Futures are shown below on Table 1.

1. Table: Accuracy results on Crude Futures forecast

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Time (weeks)** | **Open** | **High** | **Low** | **Close** |
| 2014/1 | 91,39% | 91,66% | 93,16% | 94,49% |
| 2014/2 | 94,17% | 94,07% | 90,35% | 89,68% |
| 2014/3 | 90,25% | 88,52% | 88,57% | 87,54% |
| 2014/4 | 88,03% | 89,21% | 89,18% | 90,67% |
| 2014/5 | 90,81% | 92,24% | 91,48% | 93,51% |
| 2014/6 | 93,88% | 94,46% | 97,93% | 97,99% |
| 2014/7 | 97,85% | 100,81% | 100,72% | 105,59% |
| 2014/8 | 105,77% | 103,62% | 103,88% | 102,67% |
| 2014/9 | 103,01% | 106,32% | 107,06% | 107,10% |
| 2014/10 | 108,04% | 107,61% | 108,34% | 108,03% |
| 2014/11 | 109,36% | 111,18% | 109,24% | 110,88% |
| 2014/12 | 111,08% | 104,76% | 105,27% | 100,78% |
| 2014/13 | 101,32% | 101,96% | 100,96% | 103,66% |
| 2014/14 | 103,76% | 103,85% | 103,54% | 103,54% |
| 2014/15 | 103,77% | 102,34% | 101,49% | 101,92% |
| 2014/16 | 101,83% | 105,33% | 108,17% | 112,38% |
| 2014/17 | 112,01% | 112,91% | 114,36% | 113,67% |
| 2014/18 | 114,52% | 111,66% | 113,25% | 109,12% |
| 2014/19 | 109,62% | 107,19% | 108,55% | 106,07% |
| 2014/20 | 106,16% | 105,08% | 107,42% | 105,43% |
| 2014/21 | 105,84% | 105,20% | 107,67% | 105,16% |
| 2014/22 | 105,27% | 106,64% | 107,42% | 108,24% |
| 2014/23 | 108,43% | 104,33% | 107,50% | 103,80% |
| 2014/24 | 104,18% | 103,08% | 104,98% | 103,35% |
| 2014/25 | 103,39% | 106,73% | 107,74% | 108,52% |

Not only accuracy, but directional stability is an important metrics. The related results are shown on Table 2.

2. Table: Directional stability and correlation of Crude Futures forecast

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | Open | High | Low | Close |
| Directional stability | 41,67% | 50,00% | 54,17% | 45,83% |
| Correlation | 0,83 | 0,85 | 0,90 | 0,85 |

The results of directional stability around 50% - as no real value addition - indicated the model is not as universal as it was hoped. Conclusion: other or new approach was needed to pass the stock market test.

**Theory**

Similarity analysis is used by My-X Research Team and had proved its universality and multidimensional aspects several times. The long term goal is to create a forecast model that can compete with it. For the first step a single dimensional model was created that uses similarity searching tool available in MS Excel.

Similarity search is also called as proximity search or nearest neighbor search (NNS). The very baseline of this methodology is to identify a starting point and measure the distances of its surroundings. The shorter the distance, the greater the similarity between the two points.

When time series are forecasted it is better to forecast more than one period. Thus similarity search should be continuous. Yufei et al. drew up the same problem as “A continuous nearest neighbor query retrieves the nearest neighbor (NN) of every point on a line segment (e.g., ‘find all my nearest gas stations during my route from point s to point e’).” During time series forecast this route is basically unknown, but it can be estimated in a linear approach. The next problem is the length of this line. To solve the length problem the similarity search model uses points from the time series, limited up to five to determine the direction of them.

**Method**

The first step in similarity search model is to measure the distances of the points of the time series by dividing them as chain ratios. The distance between the minimal and the maximal ratios is divided to k equal parts. Where k is an integer and cannot be more than a single digit. (Later this limitation was removed.)

Using the function CONCATENATE the previously created category numbers (e.g. 1, 2, 3, 4, 5, 6, 7, 8, 9) are merged together like a four digit number (e.g. 1234). This gives a pattern that could be searched for and on. Starting the similarity searching – according to VLOOKUP logic - with the last known pattern (n) and looking for its most similar pair (m). As it was mentioned before the nearest neighbor method will result the other closest point of the origin. So in that case, if 1234 is searched for than e.g. 1235 can be the closest match. The VLOOKUP sense them as four digit numbers and works like it was described previously.

For the end of this section the present – the last known sequence – is identified similarly to one of the past sequences. Because the goal of the model is to predict future categories, if a similar past is identified then it is easy to determine possible next category, which is given as the first element of the next sequence (m+1). This will be the first forecasted category number. The next search starts with sequence m+1, as we predicted it as the following of n. It goes repeatedly until the last forecast needed period.

When the forecast is completed and the relevant new data are given, according to the category accuracy and directional stability the forecasted values are evaluated. The schematic structure of the similarity searching model is shown on Figure 2.

 0. Times series

1. tn/tn+1; tn+1/tn+2… ratio

2. Generation of non-overlapping categories

3. Generation of sequences (5 category points at most)

4. Searching for the most similar pair (m) of the last known sequence (n), and repeated search for the sequence m+1 with its first element is stored

5. First elements of the sequences m+1, 2 ,3…z are stored until z

6. Evaluating to the actual data by category accuracy and directional stability

2. Figure: Illustration of the similarity search model

**Findings**

On the very same stock market dataset focused only on the OPEN values, the similarity search model with 184 weeks used to forecast the next 25 achieved 72% of directional stability, with 7 categories took into account, 11 cases were direct hit (44%), 11 cases were +/-1 category difference (44%), and 1 case +/-2 category differences (4%). The step between the categories was 0,0312 point. According to these achievements it seemed more successful as the classical forecast model was.

3. Table: Similarity searching model results on Crude Futures

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **Category difference** | **0** | **1** | **2** | **3** | **4** | **5** | **6** |
| **25 cases** | 11 | 11 | 1 | 1 | 1 | 0 | 0 |
| **100,00%** | 44,00% | 44,00% | 4,00% | 4,00% | 4,00% | 0,00% | 0,00% |
| **Directional accuracy:** | 18 |   |
| **Directional stability:** | 72,00% |   |

A special case emerged during testing. What if there is no similar pair of the last known sequence? This could mean there is a new pattern developing in the present without any previous indication. To bypass this case a new function was implemented in the model. Using ‘IFERROR then equal to a given value’ with VLOOKUP will result to let solution of the VLOOKUP go on, but if it comes up with an error (e.g. #MISSING), then it will result to preliminary value, which is the first sequence. By that way the model can calculate forward, but the bypass itself most likely will be a missed value.

Further tests were made using the similarity searching model to identify its performance characteristics. The next test was very similar to the Crude Futures one. Testing it on Apple share list (AAPL) with 184 weeks used to forecast the next 25 achieved 72% of directional stability, with 7 categories took into account, 7 cases were direct hit (28%), 11 cases were +/-1 category difference (44%), and 5 cases +/-2 category differences (20%). The step between the categories was 0,0349 point.

4. Table: Similarity searching model results on AAPL

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **Category difference** | **0** | **1** | **2** | **3** | **4** | **5** | **6** |
| **25 cases** | 7 | 11 | 5 | 2 | 0 | 0 | 0 |
| **100,00%** | 28,00% | 44,00% | 20,00% | 8,00% | 0,00% | 0,00% | 0,00% |
| **Directional accuracy:** | 18 |   |
| **Directional stability:** | 72,00% |   |

The Crude Futures and AAPL showed as it was expected close results. The same performance of directional stability (72%), but the AAPL was not as accurate as the Crude Futures. The next examination was different in the way of time unit. Previously the time series were in week units, during the NASDAQ test the time unit was in days. With 290 minutes used to forecast the next 194 achieved 73.2% of directional stability, with 7 categories took into account, 55 cases were direct hit (28,35%), 86 cases were +/-1 category difference (44,33%), and 36 cases +/-2 category differences (18,46%). The step between the categories was 0,0000982 point.

5. Table: Similarity searching model results on NASDAQ

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **Category difference** | **0** | **1** | **2** | **3** | **4** | **5** | **6** |
| **194 cases** | 55 | 86 | 36 | 15 | 2 | 0 | 0 |
| **100,00%** | 28,35% | 44,33% | 18,46% | 7,73% | 1,03% | 0,00% | 0,00% |
| **Directional accuracy:** | 142 |   |
| **Directional stability:** | 73,00% |   |

Comparing the newest results to last ones, it appears the model has stability. Both the directional stabilities and overall category accuracies are close to each other. For the next test phase of the model it should work in a different environment. From stock market to product sales time series of an industrial company, but with time units of weeks. It is the same dataset the classical forecasting model was represented. The similarity searching model with 100 weeks used to forecast the next 50 achieved 74% of directional stability, with 7 categories took into account, 24 cases were direct hit (48%), 17 cases were +/-1 category difference (34%), and 4 cases +/-2 category differences (8%). The step between the categories was 0,8276 point.

6. Table: Similarity searching model results on sakes time series of one product

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **Category difference** | **0** | **1** | **2** | **3** | **4** | **5** | **6** |
| **50 cases** | 24 | 17 | 4 | 2 | 1 | 2 | 0 |
| **100,00%** | 48,00% | 34,00% | 8,00% | 4,00% | 2,00% | 4,00% | 0,00% |
| **Directional accuracy:** | 37 |  |
| **Directional stability:** | 74,00% |  |

According to these results the model shows stability in a different environment from its original one. The final phase of this testing procedure is the test with random numbers. They were generated by the RANDBETWEEN function of MS Excel. After 1500 model ran, the average results are with 100 random numbers used to forecast the next 50 achieved 65,61% of directional stability, with 7 categories took into account, 14,19% of the cases had direct hit, 22,92% of the cases had +/-1 category difference, and 19,34% of the cases had +/-2 category differences. Taking into account during the random number test there is no linear approach or trend line which exists long enough the model can rely on.

7. Table: Similarity searching model results on random test

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **Category difference** | **0** | **1** | **2** | **3** | **4** | **5** | **6** |
| **100,00%** | 14,18% | 22,92% | 19,34% | 15,76% | 12,79% | 9,30% | 5,36% |
| **Directional stability:** | 65,61% |  |

Pattern recognition was also found with the model. It means, according to the serial numbers which were attached to every four digit sequence, the similarity searching model is able to recognize and follow patterns which are similar to the starting sequence, and able to switch on to another pattern without any manual interference. This kind of pattern recognition was unavailable under the random number testing due to its nature.

According to mathematical probability in a range of 1-7 numbers (where all are integers), the chance to hit that the directly upcoming – short term – random number will be greater or less than the present one is 67,34%. On the long term, this chance decreases to 36,73%. As it was presented previously the average directional stability of the similarity searching model on random numbers are 1,73% less accurate as the mathematical maximum of short term. But significantly higher as the long term one and the model did it 50 periods long.

**Conclusions**

The similarity searching model proved itself as a more universal forecasting model as the classical one. Its results showed operational stability and notable directional stability on every test so far achieved. Especially the random testing raised an interesting question, what would happen if a forecasting model can reach and later on exceed the maximum of mathematical probability on random numbers?

The next main phase of the model is to form it multidimensional. For this task, independent and depended datasets are need (e.g. classical production function). The Euclidean algorithm might be a possible solution, as it compares its points to each other. Andoni (2009) mentions it as an existing solution of nearest neighbor problems. The hybridization of the VLOOKUP logic and Euclidean algorithm could have mutually reinforcing effect.

**References**

Andoni, Alexandr: Nearest Neighbor Search: the Old, the New, and the Impossible (2009), PhD dissertation, Massachusetts Institute of Technology, <http://www.mit.edu/~andoni/thesis/main.pdf>, time of download 2015

Pitlik, László: StockNet <http://miau.gau.hu/miau/08/snbook.doc>

Pitlik, László: Online Tool for Similarity Analysis (2012), <http://miau.gau.hu/myx-free/>

Szigeti, Gábor et al.: The maximum available 'goodness' of a time series forecast (2015)

Tao, Yufei et al.: Continuous Nearest Neighbor Search (2002), Proceeding VLDB '02 Proceedings of the 28th international conference on Very Large Data Bases Pages 287-298, <http://www.vldb.org/conf/2002/S09P02.pdf>, time of download: 2015

Zoltán, Varga: A special solution for sales forecasting problems (2013), <http://miau.gau.hu/miau/178/triiilak.doc>, time of download 2015