Bayesian and Graph Theory Approaches to Develop Strategic Early Warning Systems for the Milk Market

Furkan Gürpınar[1,⋆], Christophe Bisson[2], and Öznur Yaşar Diner[2]

[1] Boğaziçi University, Istanbul, Turkey
furkan.gurpinar@boun.edu.tr
[2] Kadir Has University, Istanbul, Turkey
{cbisson,oznur.yasar}@khas.edu

Abstract. This paper presents frameworks for developing a Strategic Early Warning System allowing the estimatation of the future state of the milk market. Thus, this research is in line with the recent call from the EU commission for tools which help to better address such a highly volatile market. We applied different multivariate time series regression and Bayesian networks on a pre-determined map of relations between macro economic indicators. The evaluation of our findings with root mean square error (RMSE) performance score enhances the robustness of the prediction model constructed. Finally, we construct a graph to represent the major factors that effect the milk industry and their relationships. We use graph theoretical analysis to give several network measures for this social network; such as centrality and density.

Keywords: Strategic Early Warning System, Bayesian networks, Graph theory, forecasting, Milk.

1 Introduction

Due to the growing complexity and uncertainty of the economy, it has been underlined that current strategic decision tools are limited [1] and allow at best a reaction to threats or opportunities [2]. Thus, in using traditional strategic tools, managers are informed too late, and too often decisions are made on the basis of heuristics [3]. When trying to overcome these limitations and strengthen strategic planning and governance,the importance of Strategic Early Warning Systems' (SEWS)has been raised [4]. SEWS can help decision-makers anticipate market changes.

SEWS integrate scenario techniques which aim to create alternative pictures of the future and to challenge mental models [5]. The general framework [2] of SEWS is: 1) Determine drivers of change; 2) generate scenarios; 3) Make a strategic simulation for each scenario (named War Game); 4) Implement the system by watching the drivers of change (Use of Competitive Intelligence methods and tools for this) which could lead to the appearance of a pre-determined scenario then launch an alert to anticipate either a threat or opportunity. Our research focuses on the first two steps of the framework.

⋆ This project is funded by Kadir Has University BAP 2014-07.

Á. Rocha et al. (eds.), *New Contributions in Information Systems and Technologies*,
Advances in Intelligent Systems and Computing 353, DOI: 10.1007/978-3-319-16486-1_52

Although several qualitative methods of SEWS were developed which demonstrated their importance for governance [6], no quantitative method of SEWS has been developed [4]. Indeed, strategy deals with a high number of variables and the use of qualitative methods can be deemed as adapted, especially when the model includes human behavior as one can construe this variable as difficult to predict. Thus, we aim to address an important scientific gap by applying different multivariate time series regression and Bayesian networks on the two first steps of the general frame of SEWS to predict the impacting scenario. By applying graph theory on the same two steps of SEWS, it limits the number of possibilities of a strategic system which can be a useful tool for better estimation of scenario.

We chose the milk sector in France for our experiment in line with the call from the EU Commission [7] for more robust tools to better predict the milk price and anticipate changes on this market. Indeed, the milk price is highly volatile. For instance, French farmers income can vary by over one third from one year to the next [8]. For example, 1% or 2% discrepancy between supply and demand can trigger a variation of 50% to 100% in income [8]. Yet, the new Common Agricultural Policy which will go into effect in 2015, will end quotas for the milk. Hence, there is currently high uncertainty in the EU market regarding the consequences of the end of quotas on the milk price.

In Section 2 , we build the Bayesian model; apply it on our data and discuss about the results obtained through Bayesian analysis. Then, we construct a graph to represent the milk network and give several relationship measures. Finally, we discuss our findings and conclude with comments on future work.

2 Theoretical Background and Methodology

Concerning the application of graph theory, we collaborated with a French milk expert to provide the input (e.g. impacts and probabilities of impacts) to evaluate the environment for milk and determine the drivers of change [2]. We used 7 forces [2] to evaluate the micro environment; indicators based on the 5 forces of Porter to which were added two new forces. Thereby, highlighting the rivalry between established firms, the barriers to enter the market, the products / services / technologies of substitution, the bargaining power of customers and the bargaining power of suppliers [9] to which were added two new forces i.e. the bargaining power of employees and complementary products / services / technologies. To understand the macro environmental changes, the well known PESTEL analysis [10] was used which covers the political, economic, social, technological, environmental and legislative frameworks.

The review of recent literature for the milk market underlines that only the barriers to enter the market are not variables for milk in the coming years (and confirmed by the French milk expert) as they will continue to be high [11]. Therefore, we obtained a 12-by-12 matrix by adding the 6 drivers of change from the micro environment and 6 from the macro environment. Thereafter, graph theory was applied to the evaluated change drivers to enhance the major factors that affect the milk industry and their relationships.

To establish a broader understanding, we first present our work in the form of a work flow diagram, as shown in Figure 1 below.

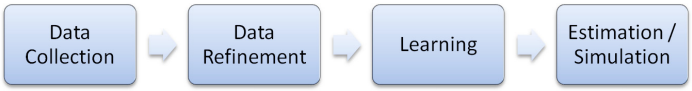

Fig. 1. System Pipeline

2.1 Data Collection

Data for Bayesian Approach. After determining the important parameters, we started the quantitative analysis by collecting time series data for various macro-economic indicators related to milk, which are the world milk demand and production, the consumer price index for milk-related products, livestock and input costs (e.g. energy). We collected time series data for the period from January 1990 to October 2014. Annotating the time $t = 0$ at the beginning of our observations, we have $T = 295$ time points where observations are recorded. The time series data can be found at the website of INSEE (the French official statistic public organisation), an example data link is in [12]

Data for Graph Analysis. We collected data from one French milk expert, to assess the relationships between the 12 variables, as defined in Section 1. We prepared questionnaires that ask for the impact and their probability. In other words, we obtained 2 matrices of size 12 by 12 which contain 6 points Likert-scale evaluations, where each cell is an integer between 0 and 5.

2.2 Data Restoration

Since the time series data for various indicators mentioned in Section 2.1 came from different sources, some of them are measured in different units of time, such as monthly, quarterly, yearly. Therefore to establish a consistent data set, we used various mathematical methods to convert all the time series to monthly-observed variables. In that aim, we chose the Least-Squares approximation method for interpolating and extrapolating the data and therefore to estimate the best possible values for intermediate time steps.

Least-Squares Fitting. The aim of the Least-Squares Method (LSM) is to establish a linear model constructed depending on the observation pairs $[x_i, y_i]$. Therefore we tried to obtain the best possible prediction line $y' = \beta_0 + \beta_1 x'$. The aim here was to estimate the best values for β_0 and β_1. To achieve this we needed to solve the following equation:

$$\beta = \arg\min_{b \in \mathbb{R}^p} S(b) = \left(\frac{1}{n} \sum_{i=1}^{n} x_i x_i' \right)^{-1} \cdot \frac{1}{n} \sum_{i=1}^{n} x_i y_i \tag{1}$$

Or in linear form :

$$\beta = (X^T X)^{-1} X^T y \, . \tag{2}$$

Using the set of observations (X : input variables) and (y : output variables) we obtained a solution which minimizes the sum of squared errors.

2.3 Discretizing the Data

We simplified the learning problem here by converting the time series signals to discrete values, then any given signal x is transformed to $f(x) = x_d$. In the new form x_d, every element x_{di} could have a value between 1 and V, where V is the cardinality of the state space. So intuitively, the values of x_d actually represented changes in the data (1: Big drop, 2: Smaller drop , ... V : Big rise) For consistency with the Likert scale data, we usually set V=5 but we also saw that increasing V increased the signal reconstruction accuracy of our system.

Discretization Parameters. Since we converted our continuous signals to discrete ones, we had to assess V different states to the elements of x_d, in terms of numerical changes. Furthermore, as the range of various indicators changed greatly (e.g. from the range of 10^{-2} to 10^{5}, we were not able to use a fixed change index. Therefore, we needed to establish V different amounts of change for each signal, depending on observed values. To find a reasonable set of changes, we used the K-means clustering algorithm as described in the following section.

K-means Clustering. K-means clustering algorithm is a way to perform vector quantization, by finding optimal set of clusters, and assigning each sample in the vector to a cluster center. [13]
In our application, we converted the signal to a format x_c where it represented changes in the data, so defining it formally : $x_{ci} = x_i - x_{i-1}$. Therefore we apply k-means clustering method on this change vector x_c, to find the V most observed change values in the samples, and assigning each sample to one of the V cluster centers, we have the discrete vector x_d, as mentioned in Section 2.3

2.4 Learning the Probability Distributions

Since our aim was to estimate the future values of dependent variables, we first needed to obtain prior probabilities to feed our Bayesian decision system. For this, we used two different probability definitions, which can then be combined in a single set of matrices. Two different probability estimations are explained below.

Intra-variable (Temporal) Probabilities. We started by finding the probability distributions of single variables over different time lags. In other words, we constructed probability distribution function (PDF) tables to establish the prior probability of observing variable i having the value $k_1 \in [1 : V]$ when observed that it has the value $k_2 \in [1 : V]$, on time $(t - lag)$. Thus, we established a seasonal model where we have an estimation of probabilities of observing a single variable. After normalization, this yields a (VxV) PDF table T. Where $T_{l,i,j} = p(x_t|x_{t-l})$, in other words the probability of observing $x = j$ when we know that $x = i$ [lag] periods before.

Inter-variable (Transitional) Probabilities. Similarly to the intra-variable approach, we also constructed prior probabilities which represent the effect of indicator variables on the dependent variables over different time lags, more formally $p(y_t \mid x_{1,(t-l)}, x_{2,(t-l)},$
$...x_{I_y,(t-l)})$

Finally, we obtained a set of $V - by - V$ probability distribution matrices from the collected set of data. For the representation of PDFs, assuming each variable depends on each other (a complete graph), we have a data structure of size N^2 by V^2 where N is the number of variables in the model, and V is the cardinality of the state space.

2.5 Bayesian Decision Making

We used the Bayes' Formula for decision making following the first 2 steps of the SEWS framework, which is shown in Equation 3 below.

$$P(\theta|\mathbf{D}) = P(\theta)\frac{P(\mathbf{D}|\theta)}{P(\mathbf{D})} \tag{3}$$

Bayes' theorem indicates that we can estimate the probability of an observation θ, given data D, by using the probability of θ itself, the probability of having D when θ is observed, and the normalization factor which is the probability of D itself. The left hand term is called the posterior probability of θ which is computed using the prior probabilities on the right-hand side of Equation 3. We compute the prior probabilities as described in Section 2.4, and use the posteriors to evaluate scenarios, which will be described in the following section.

2.6 Scenario Evaluation

Having collected all the data, graphical model and the prior probability distributions, we used our system for simulation, to determine the probability of a scenario happening T_f time periods after the last observation. Therefore in our case, a scenario S is simply represented as an N by 1 vector where each member S_i represents the numerical value of variable i, at the time period designated by $T + T_f$. Since we cannot measure the accuracy of our system's prediction with a large T_f value, we ran some simulations on the previously-observed data, which are explained in Section 3.

2.7 Graph Analysis

In this section we represent the major factors that effect the milk industy and their relationships by a graph. By analyzing this graph we give results on the relationship measures for this social network. With this aim we considered 12 factors that have an effect on the milk industry. These factors are given as follows (together with their abbreviations): Policy (Po), Economy (Ec), Regulation(Re), Social Framework(So), Technology(Te), Environment(En), Bargaining power of Workers(Bw), Bargaining power of Suppliers(Bs), Bargaining power of Customers(Bc), Substitutes(Su), Additional Products(Ap), Competition in the milk market(Co). A discussion on these factors is given in Section 2.

We define an edge weighted directed graph $G = (V, E)$ where the vertex v_i correspond to the factor i. If any change in one factor, represented by v_i, effects another factor, represented by v_j, then we put a directed edge $e = (v_i, v_j)$ with the terminal vertex being v_j. Each edge $e_{i,j} = (v_i, v_j)$ has a weight w_{ij}. The weight w_{ij} corresponds to the impact of the factor v_i on the factor v_j where w_{ij} is from the set $\{1, 2, \ldots, 5\}$. Here a weight of 1 corresponds to a very minor effect; a weight of 5 corresponds to a very major effect and the other weights are distributed according to the level of inter depedence. In case of no effect, the weight is assumed to be zero.

In order to measure the extent to which the chosen factors for the milk market are connected among themselves, we first calculate the density D of this network given by the following formula :

$$D = \frac{\Sigma_{i=1,j=1,i\neq j}^{n} w_{ij}}{n(n-1)} \tag{4}$$

Applying this to our graph, the density is found to be 2.12. This imlpies that overall the factors are fairly evenly chosen. Thus, it shows that the factors in the graph have tight relationships. Furhtermore, it implies that a collection of these factors will have a major efect on the milk industry and that the chosen factors represent the milk industry well [14].

The major centrality measure in a directed edge weighted graph is given by nodal indegrees and nodal outdegrees. For a better analysis of the graph model, we calculated the nodal degrees using the summation method and the average method; and compared the corresponding results. According to the summation method, with a common sum of 32, Te, En, and Bc have the maximum nodal outdegree followed by Co. This shows that these three factors (Te, En, and Bc) have the largest major effect on the other factors; meaning that a change in any one of them will effect the milk industry more drastically than the other factors. In the average method one calculates the mean value of the edge weights different from zero. When the average method is used, we see that the top 3 major factors remain the same, and they are followed by Bs.

As far as nodal indegrees are concerned, according to the summation method Co has the maximum value followed by Ec. Thus the summation method captures Co as the recipient of the most attraction with varying intensities, thus, ranking Co as "the most popular" factor. On the other hand, when the average method is used Co still remains the highest but So is the second next largest factor in terms of indegrees. Therefore So has more intensive ties than Co with the other factors affecting the milk industry. Note that according to both methods, Bw has the lowest indegree and the lowest outdegree. This implies that the bargaining power of workers is weak compared to the other factors and thus, it should be strengtened in order to have a more balanced milk network.

Further analysis is made by looking at the subcliques of G. Knowing the maximum total edge weighted subcliques of the graph will give crucial information about the firmness of ties between a given subset of factors. A solution to this problem for a fixed number of factors, would reveal the location of the strength of this social network. We can use this information in the SEWs model as follows: If we detect any change in a factor whose corresponding vertex is contained in a maximal clique, then we know

that, authomatically, all factors in this clique will be effected and therefore must be emphasized in the early warning system that will be generated.

For this end, we reduce the directed graph into a nondirected graph where the edge e_{ij} will have weight $w_{ij} + w_{ji}$. By making this reduction, although we loose directional relations, it allows us to do analysis concerning weighted subcliques. Given a graph G and a clique size k it is NP-complete to find a subclique of size k and of maximum total edge weight. Thus this problem is computationally hard. For instance, the subclique formed by Po, Ec, So and En give a total weight of 39. Here we are interested in detecting, if it exitsts, a subclique of G of size 4 with total edge weight larger than 39.

3 Experimental Setup

Using the prior probabilities explained in Section 2, we used Bayes' decision theorem to forecast the future values of our time-series signals. We represented our system's state by N discrete time-series signals of length T, hence a T-by-N matrix. We fed this matrix into our simulation code and we obtained a new scenario of size $(T + 1)$ x N. The process is repeated until we reach time T_f, and converting the discrete signals back to the numerical values, we estimated the final values of variables. To analyze the probability of scenarios, we repeated this process a large number of times, hence we obtained a probability distribution function for the scenario at time T_f.

3.1 Simulation

Since our aim was to obtain a probability density function for the final values of the variables, we ran N ($\sim 10^3$) simulations to construct the future values of the discrete time series vectors, and converting them back to continuous signals, we obtained one final value per per parameter for each turn. Collecting all the final values, we obtained a data distribution. By fitting a normal distribution on this data, we obtain a probability for a given scenario.

To evaluate the accuracy of our framework, we ran some tests on different parts of the machine learning system, and we reported performance scores in the following section.

4 Results

4.1 Signal Reconstruction Accuracy

In this section we analyzed the accuracy of our signal conversion system. As explained in Section 2.3, we converted our time series data into discrete values. Hence we needed to reconstruct the signal back to a "continuous" time-series form, which inevitably caused information loss. Intuitively, increasing the number of cluster centers, k, in K-means clustering should decrease the reconstruction error. Here we present a chart for a sample signal (namely the input cost of dairy products) which shows the relation between the number of cluster centers and the root mean square error (RMSE) for signal reconstruction, in Figure 2 below.

An example signal reconstruction for 5 and 10 cluster centers are shown in Figure 3. As expected, the reconstructed signal converged to the original one as the number of cluster centers increases.

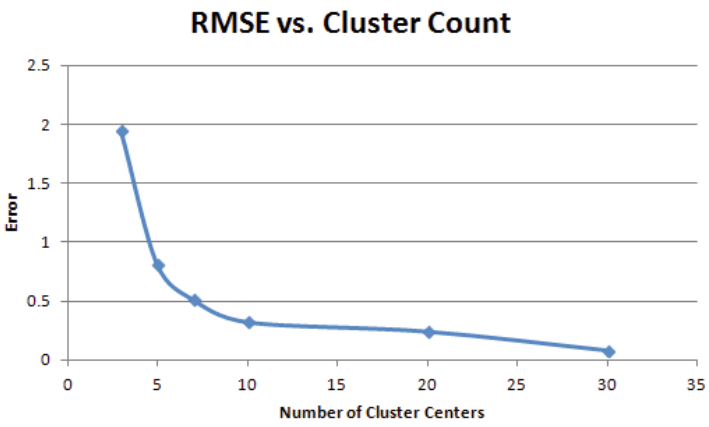

Fig. 2. Reconstruction error vs. Number of cluster centers

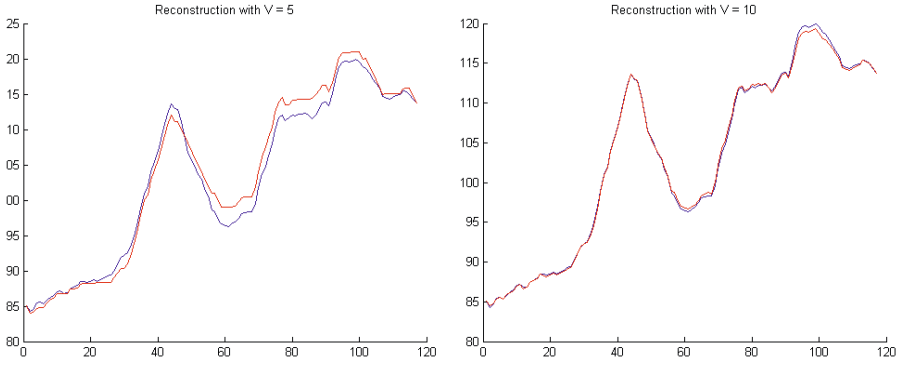

Fig. 3. Reconstruction with : (a) 5 and (b) 10 cluster centers

4.2 Forecasting Accuracy

For the evaluation of the accuracy of our prediction system, we again used the RMSE error measure, with a performance test similar to a machine learning application. In this test, we used the parameter $\tau \in [0, 1]$ which is the ratio of training set size to the data set size D. In other words, we used the first $\lfloor \tau D \rfloor$ number of observations for the learning (see Section 2.4), and we ran a simulation for the remaining $\lceil (1 - \tau)D \rceil$ "unobserved" time periods, and thus constructed a scenario which is of size D. After obtaining a large ($\sim 10^3$) number of scenarios, and taking the mean of them, we estimated the signal S' for the variable of interest. Since we already knew the original signal S, we represented our system's performance with the Root Mean Square Error $RMSE(S, S')$. Below are some results for different variables and different values of τ.

Table 1. Forecast Error vs. τ

	τ **: Training Set Ratio**					
Variable	**0.3**	**0.5**	**0.7**	**0.8**	**0.9**	**Multiplier**
Price	93.8	49.7	49.4	13.7	3.8	10^{-3}
Livestock	15.4	12.7	4.2	1.9	0.1	10^{2}
Demand	11.1	9.5	11.5	6.6	2.6	10^{0}

4.3 Scenario Probability Evaluation

In this section we used our framework to estimate the probability of different scenarios relevant to the milk market. Two scenarios were given in terms of milk price, and another one about the milk demand in the European Union [15]. We tested these scenarios by propagating the market's state upto the year 2020, with the method explained in Section 3.1. The results for the 3 scenarios are shown in Table 2 below.

Table 2. Scenario Probabilities

Scenario	**Probability**
Optimistic : Price + 15%	0.75
Pessimistic : Price - 15%	0.02
Most Likely : Demand + 2%	0.42

5 Discussion and Conclusion

We tested our algorithm with different scenarios for the variables of price and demand. As expected, the "likely" scenario resulted in a high probability value, whereas the probability of the "pessimistic" scenario (price decreases by 15%) yielded a very small probability, which can be seen from the time series where the price index clearly shows an increasing trend. This also justifies the probability of the "optimistic" scenario (price increases by 15%) being relatively high. It should also be noted that one of the scenarios of the experts [15] was given under the title of "Most Likely", which argues that the total milk demand in the EU will increase by 2%. Although its name suggested that it would yield the highest probability, our simulation assigns this a lower probability compared to the optimistic scenario.Hence we observe a difference between the prospective approach (a type of qualitative forecasting approach) and the approaches obtained with our simulation for SEWS.

We looked at the graph representing the milk industry and gave several network measures. These measures indicate an actors level of involvement in network activities and thus they give crusial information about the interpretation of the SEWs model. For future work, we plan to work on heuristic algorithms to give us an approximate anwer to the maximum weighted subclique problem that we introduced in Section 2. Additional analysis can be done by looking at the betweenness measures defined for edge weighted graphs [16] and geodesic distances [17].

If graph analysis can help to better predict the scenarios of the market, our Bayesian application on milk is purely quantitative, thus provide a numerical result. Only time

will tell which approach is more accurate but if our predictions based on Bayesian are true, one could construe this as a leapfrog in the field.

References

1. Gilad, B.: Business War Games: How large, small, and new companies can vastly improve their strategies and outmaneuver the competition. Career press (2009)
2. Bisson, C.: Guide de gestion stragique de l information pour les PME. Montmoreau. Les2Encres, France (2013)
3. Bisson, C., Guibey, I., Laurent, R., Dagron, P.: Mise en place d'un Systéme de détection de Signaux Précoces pour une Intelligence Collective de l'Agriculture appliquée aux filiéres de l'élevage bovin. Paper presented at the PSDR Symposium, Clermont Ferrand, France (2012)
4. Fuld, L.M.: The Secret Language of Competitive Intelligence: How to See Through & Stay Ahead of Business Disruptions, Distortions, Rumors & Smoke Screens. Dog Ear Publishing (2010)
5. Schwarz, J.O.: Pitfalls in implementing a strategic early warning system. Foresight 7, 22–30 (2005)
6. Gilad, B.: Early warning: using competitive intelligence to anticipate market shifts, control risk, and create powerful strategies. AMACOM Div American Mgmt Assn (2003)
7. Commision, E.: Report from the commission to the european parliament and the council. Evolution of the market situation and the consequent conditions for smoothly phasing-out the milk quota system second soft landing report. Brussels (December 10, 2010)
8. Momagri: Chiffres clés de l'agriculture (2012)
9. Porter, M.: Competitive advantage of nations (1990)
10. Kotler, P., Armstrong, G.: Marketing. Praha (2004)
11. Chevalier, F., Veyssiere, L., Buccellato, T., Jicquello, J., de Oteyza, C.: Analysis on future developments in the milk sector. Prepared for the European Commission - DG Agriculture and Rural Development (2012)
12. National Institute of Statistics and Economic Studies: Macro-economic database (2014), `http://www.bdm.insee.fr/bdm2/choixCriteres?codeGroupe=1466` (accessed November 18, 2014)
13. MacQueen, J., et al.: Some methods for classification and analysis of multivariate observations. In: Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability, California, USA, vol. 1, pp. 281–297 (1967)
14. Wasserman, S.: Social network analysis: Methods and applications, vol. 8. Cambridge university press (1994)
15. Pôle régional Économie & Prospective: Quel élevages laitiers en normandie 2020. Synthèse (2014)
16. Freeman, L.C., Borgatti, S.P., White, D.R.: Centrality in valued graphs: A measure of betweenness based on network flow. Social Networks 13, 141–154 (1991)
17. Yang, S., Knoke, D.: Optimal connections: strength and distance in valued graphs. Social Networks 23, 285–295 (2001)