Research Article

Semiconductor Yield Forecasting Using Quadratic-Programming-Based Fuzzy Collaborative Intelligence Approach

Toly Chen and Yu-Cheng Wang

Department of Industrial Engineering and Systems Management, Feng Chia University, Taichung City 407, Taiwan

Correspondence should be addressed to Toly Chen; tcchen@fcu.edu.tw

Received 26 March 2013; Accepted 17 May 2013

Academic Editor: Yi-Chi Wang

Several recent studies have proposed fuzzy collaborative forecasting methods for semiconductor yield forecasting. These methods establish nonlinear programming (NLP) models to consider the opinions of experts and generate fuzzy yield forecasts. Such a practice cannot distinguish between the different expert opinions and cannot easily find the global optimal solution. In order to solve some problems and to improve the performance of semiconductor yield forecasting, this study proposes a quadratic-programming- (QP-) based fuzzy collaborative intelligence approach.

1. Introduction

Yield is the proportion between goods and total products. Given the same inputs, the higher yield means more salable outputs that increase the revenues. Yield is a key performance measure in operations and has a dominant effect on manufacturing economics. Yield improvement is also a crucial work for a factory's competitiveness [1]. In addition, predicting the future yield to estimate the possible gains is also a basis for long-term production planning [2]. The ramping of a semiconductor manufacturing factory should occur quickly and then product yields should be maintained at a high level to maximize profits. This study is therefore committed to predicting the yield of a semiconductor product. Semiconductor products are focused on because of the extensive use in modern life, such as personal computer, laptop, cell phone, and others. According to Chen and Wang [3], methods for predicting the yield of a semiconductor product can be divided into two categories—macro yield modeling (MaYM) methods and micro yield modeling (MiYM) methods. This study belongs to the MaYM category, in which a learning model is used to predict the future yield of a semiconductor product. After conversion, some yield learning model can be solved using linear regression [3], which is convenient and explains why the related methods have been welcomed. However, there is much uncertainty in the yield learning process. To deal with this issue, several treatments have been taken in the literature (Figure 1).

(1) Probabilistic/Stochastic Methods. Methods of this sub-category assume that parameter distributions are known in advance to a certain degree, and these distributions can be modified in a Bayesian manner after actual values are observed. The works of Spence [4], Majd and Pindyck [5], Mazzola and McCardle [6], and Anderson [7] all belong to this sub-category.

(2) Fuzzy Methods. Methods of this sub-category fit the yield learning process with a possibility regression model, and yield forecasts are given in fuzzy values to consider the uncertainty [1, 3, 8–12]. In the work of Chen and Wang [3], the fuzzy yield learning curve was fitted by solving a linear programming (LP) problem. Wu et al. [12] established a fuzzy back propagation network (FBPN) to predict the yield of semiconductor. The inputs of this network include three categories: the physical parameters, electrical test parameters, and wafer defect parameters. Chiang and Hsieh [13] applied a similar fuzzy neural network...
method, in which grey relational analysis was used to find the most influential parameters. Lin [11] considered the uncertainty of the defect clustering pattern and established a number of fuzzy rules to predict yield.

On the other hand, the concept of fuzzy collaborative intelligence was proposed by Pedrycz in 2003 [14] and has been applied to the forecasting of various properties, such as yield [8, 9], global CO$_2$ concentration [15], job cycle time [16], and others. Among them, yield forecasting received much attention and is the focus of this study. In [8], Chen and Lin proposed a fuzzy collaborative intelligence approach for forecasting the yield of a semiconductor product. In the fuzzy collaborative intelligence approach, a group of experts predict the future yield with a fuzzy number from different points of view. The forecasts by the experts are aggregated in such a way that both the precision and accuracy can be improved at the same time.

1. Accuracy: the forecasted value should be as close as possible to the actual value.

2. Precision: a narrow interval containing the actual value is established.

However, the fuzzy collaborative intelligence approach requires the solving of two nonlinear programming (NLP) problems, which is not easy. Furthermore, the existing optimization packages may not be able to guarantee the global optimality of the solutions. In addition, the views of the experts are not reflected properly on the objective function. In order to solve these two problems, the following treatments have been taken in this study.

1. A new partitioning method has been proposed, which considers the effects on the objective function in partitioning the range of parameters.

2. The nonlinear objective functions and constraints were converted into quadratic ones using several polynomial fitting techniques. Then, a variety of methods, such as the interior point method, the active set method, the augmented Lagrangian method, the conjugate gradient method, the gradient projection method, and the extension of the simplex algorithm, can be used to solve the quadratic programming (QP) problems.

The differences between the proposed methodology and the previous methods are summarized in Table 1.

The rest of this paper is organized as follows. Section 2 reviews the related literature. In Section 3, we propose some countermeasures to solve the problems in the work of Chen and Lin [8], in order to propose a QP-based fuzzy collaborative forecasting method. Examples in the work of Chen and Lin [8] were used to assess the effectiveness of the proposed methodology, and to make a comparison with the NLP-based method. Finally, the conclusions of this study are made in Section 4.

2. Literature Review

Multiple analyses of a problem from diverse perspectives raise the chance that no relevant aspects of the problem will be ignored. In addition, as internet applications become widespread, dealing with disparate data sources is becoming more and more popular. Technical constraints, security issues, and privacy considerations often limit access to some sources. Therefore, the concepts of collaborative computing intelligence and collaborative fuzzy modeling have been proposed, and the certain so-called fuzzy collaborative systems are being established. In a fuzzy collaborative system, some experts, agents, or systems with various backgrounds are trying to achieve a common target. Since they have different knowledge and points of view, they may use various methods to model, identify, or control the common target. The key of such a system is that these experts, agents, or systems share and exchange their observations, settings, experiences, and knowledge with each other when achieving the common goal. This features the fuzzy collaborative system distinct from the ensemble of multiple fuzzy systems.

In the limited literature, fuzzy collaborative intelligence and systems have been successfully applied to collaborative clustering, group forecasting, agent negotiation, assessment and reputation, information filtering, robot control, intrusion detection, object tracking, and so forth. Applications of fuzzy collaborative intelligence in other fields remain to be investigated. Several studies have argued that for certain problems, a fuzzy collaborative intelligence approach is more precise, accurate, efficient, safe, and private than typical approaches.

Designing is a work in which the views may vary and therefore needs coordination or cooperation. Shai and Reich [17, 18] defined the concept of infused design as an approach for establishing effective collaboration between designers from different engineering fields. Ostrosi et al. [19] defined the concept of consensus as the overlapping of design clusters of different perspectives.

In demand forecasting, Lo et al. [20] believed that the demand plan has to be compromised and achieved through the collaborative efforts of all the demand network stakeholders. Büyükoğlan and Vardaloğlu [21, 22] applied the fuzzy cognitive map method to the collaborative planning, forecasting, and replenishment of a supply chain. The initial
Table 1: The differences between the proposed methodology and the previous methods.

<table>
<thead>
<tr>
<th>Method</th>
<th>Expert collaboration</th>
<th>Partitioning the ranges of parameters</th>
<th>Models to solve</th>
<th>Easiness to solve</th>
<th>Optimality of the solution</th>
</tr>
</thead>
<tbody>
<tr>
<td>Chen and Wang [9]</td>
<td>No</td>
<td>No</td>
<td>LP</td>
<td>Very easy</td>
<td>Global optimal</td>
</tr>
<tr>
<td>Chen and Lin [8]</td>
<td>Yes</td>
<td>Uniform partitioning</td>
<td>NLP</td>
<td>Very difficult</td>
<td>Local optimal</td>
</tr>
<tr>
<td>Wu et al. [12]</td>
<td>No</td>
<td>No</td>
<td>FBPN</td>
<td>Very difficult</td>
<td>Local optimal</td>
</tr>
<tr>
<td>The proposed methodology</td>
<td>Yes</td>
<td>Considering the effects on the objective function</td>
<td>QP</td>
<td>Easy</td>
<td>Global optimal</td>
</tr>
</tbody>
</table>

values of the concepts and the connection weights of the fuzzy cognitive map are dependent on the subjective belief of the expert and can be modified after collaboration. Cheikhrouhou et al. [23] thought that collaboration is necessary because of the unexpected events that may occur in the future demand.

In the collaborative forecasting, according to Poler et al. [24], the comparison of collaboration methods and the proposing of software tools, especially regard forecasting methods for collaborative forecasting, are still lacking. Chen [16] put forward a fuzzy collaborative forecasting method to predict the job cycle time in a wafer fabrication factory, in which each expert used a fuzzy multiple linear regression equation to predict the cycle time of a job. The cycle time forecasts of different experts were aggregated using the hybrid fuzzy intersection (FI) and back propagation network (BPN) approach. A similar method has been used by Chen and Wang [15] to predict the global CO\textsubscript{2} concentration. The traditional method in this field is a simple time series analysis, but it is not accurate enough. Nonlinear methods, such as artificial neural networks [25, 26], may be more effective.

In all fuzzy collaborative intelligence methods, the consensus of results is being sought. Pedrycz and Rai [27] discussed the problem of collaborative data analysis by a group of agents having access to different parts of data and exchanging findings through their collaboration. A two-phase optimization procedure was established, so that the results of communication can be embedded into the local optimization results. Sometimes, however, the consensus may not exist. Therefore, Chen [28] defined the concept of partial consensus as the intersection of the views of some experts. In the work of Chen and Wang [9], an agent considered the settings of other agents, to optimize its own setting.

### 3. QP-Based Fuzzy Collaborative Intelligence Approach

According to Gruber [29], the general yield learning model of a semiconductor product is

\[ Y_t = Y_0 e^{-b/t + r(t)} , \]

where \( b \) and \( Y_0 \) indicate the learning constant and the asymptotic yield, respectively. \( b \geq 0; 0 \leq Y_0, Y_t \leq 1; t \) is time, \( t = 1 \sim T; r(t) \) is a homoscedastic, serially noncorrelated error term. After converting to logarithms,

\[ \ln Y_t = \ln Y_0 - \frac{b}{t} + r(t) = a - \frac{b}{t} - r(t) , \]

where \( a = \ln Y_0 \). In Chen and Lin’s fuzzy collaborative intelligence approach [8], to consider the uncertainty in yield learning, the parameters in (2) are given in triangular fuzzy numbers as follows:

\[ \hat{Y}_t = (Y_{t1}, Y_{t2}, Y_{t3}) , \]
\[ \hat{Y}_0 = (Y_{01}, Y_{02}, Y_{03}) , \]
\[ \hat{b} = (b_1, b_2, b_3) , \]
\[ \hat{a} = (a_1, a_2, a_3) , \]

and the following two NLP problems are to be solved.

NLP problem I is

\[ \text{Min } Z_1 = \sum_{t=1}^{T} (\ln Y_{t3} - \ln Y_{t1})^{q_k} \]

subject to

\[ \ln y_t \geq \ln Y_{t1} + s_k (\ln Y_{t2} - \ln Y_{t1}) , \]
\[ \ln y_t \leq \ln Y_{t3} + s_k (\ln Y_{t2} - \ln Y_{t3}) , \]
\[ \ln Y_{t1} = a_1 - \frac{b}{t} , \]
\[ \ln Y_{t2} = a_2 - \frac{b}{t} , \]
\[ \ln Y_{t3} = a_3 - \frac{b}{t} , \]
\[ 0 \leq a_1 \leq a_2 \leq a_3 , \]
\[ 0 \leq b_1 \leq b_2 \leq b_3 , \]
\[ t = 1 \sim T . \]

NLP problem II is

\[ \text{Max } Z_2 = m_k \sqrt{\sum_{t=1}^{T} s_i m_i} / T \]
subject to
\[
\sum_{t=1}^{T} \ln (Y_{t3} - Y_{t1})^{o_k} \leq T \cdot d_k^{o_k},
\]
\[
\ln y_t \geq \ln Y_{t1} + s_t \ln (Y_{t2} - Y_{t1}),
\]
\[
\ln y_t \leq \ln Y_{t3} + s_t \ln (Y_{t2} - Y_{t3}),
\]
\[
\ln Y_{t1} = a_1 - \frac{b_1}{t},
\]
\[
\ln Y_{t2} = a_2 - \frac{b_2}{t},
\]
\[
\ln Y_{t3} = a_3 - \frac{b_3}{t},
\]
\[
0 \leq a_1 \leq a_2 \leq a_3,
\]
\[
0 \leq b_1 \leq b_2 \leq b_3,
\]
\[
t = 1 \sim T,
\]

where \( \vec{b} = (b_1, b_2, b_3) \) and \( \vec{Y}_0 = (Y_{01}, Y_{02}, Y_{03}) \) indicate the learning constant and the asymptotic yield. \( t \) is time, \( t = 1 \sim T \); \( o_k \) indicates the sensitivity of expert \( k \) to the uncertainty in the fuzzy forecast; \( o_k \geq 0 \). Chen and Wang [9] suggested that \( o_k \) should be between 1 and 4. \( s_k \) is the satisfaction level required by expert \( k \); \( 0 \leq s_k \leq 1 \); \( m_k \) reflects the importance of the outlier to expert \( k \); \( m_k \geq 0 \). Chen and Wang [9] suggested that \( m_k \) should be between 1 and 4. \( d_k \) is the required range by expert \( k \), and \( d_k \geq 0 \). Maximizing \( Z_2 \) is equivalent to
\[
\text{Max } Z_2' = \sum_{t=1}^{T} e^{m_k t}.
\]

However, there are two problems with Chen and Lin's method [8].

(1) The range of \( o_k \) was divided into six segments of the same width, and seven values of \( o_k \), 1, 1.5, 2, 2.5, 3, 3.5, and 4, corresponding to seven linguistic terms (extremely insensitive, very insensitive, somewhat insensitive, moderate, somewhat sensitive, very sensitive, extremely sensitive) were provided for each expert to choose from. However, the effects of such a uniform partitioning were not proportionally reflected on the objective function. For example, the distance between 3.5 and 4 is the same as that between 1 and 1.5. If \( \ln Y_{t3} - \ln Y_{t1} = 0.5 \), then \( \ln (Y_{t3} - Y_{t1})^4 \) = \( \ln (Y_{t3} - Y_{t1})^{1.5} \) = \( \ln (Y_{t3} - Y_{t1})^{1} \) = \( -0.146 \). Obviously, \( -0.026 \gg -0.146 \).

(2) The nonlinear objective functions and constraints are difficult to handle.

First, the range of \( o_k \) should be partitioned in such a way that the range of \( \ln (Y_{t3} - Y_{t1})^{o_k} \) can be divided into six equal parts. According to (7), \( \ln Y_{t3} - \ln Y_{t1} \) is to be smaller than \( d_k \). Therefore, the problem becomes dividing the interval \( [d_k^1 d_k^1] \) into six equal parts indicated with \( o_k(j) o_k(j + 1), j = 1 \sim 6; o_k(1) = 1, o_k(7) = 4 \). To this end, the following equations are to be solved individually:
\[
d_k^{o_k(2)} = \frac{5d_k + d_k^3}{6} \rightarrow o_k(2) = \ln \left( \frac{(5d_k + d_k^3)/6}{\ln (d_k)} \right),
\]
\[
d_k^{o_k(3)} = \frac{4d_k + 2d_k^4}{6} \rightarrow o_k(3) = \ln \left( \frac{(4d_k + 2d_k^4)/6}{\ln (d_k)} \right),
\]
\[
d_k^{o_k(4)} = \frac{3d_k + 3d_k^5}{6} \rightarrow o_k(4) = \ln \left( \frac{(3d_k + 3d_k^5)/6}{\ln (d_k)} \right),
\]
\[
d_k^{o_k(5)} = \frac{4d_k + 2d_k^6}{6} \rightarrow o_k(5) = \ln \left( \frac{(2d_k + 4d_k^6)/6}{\ln (d_k)} \right),
\]
\[
d_k^{o_k(6)} = \frac{d_k + 5d_k^7}{6} \rightarrow o_k(6) = \ln \left( \frac{(d_k + 5d_k^7)/6}{\ln (d_k)} \right).
\]

For example, if \( d_k = 0.5 \), then \( o_k(1) = 1, o_k(2) = 1.22, o_k(3) = 1.5, o_k(4) = 1.83, o_k(5) = 2.26, o_k(6) = 2.88, \) and \( o_k(7) = 4 \). Subsequently, the nonlinear objective function and constraints can be converted into quadratic ones by minimizing the mean absolute error of approximation when \( 0 \leq x \leq 1.2 \) as
\[
\text{Min } \sum_{x=0.01}^{1.2} \left| x^k - (\alpha x^2 + \beta x + \gamma) \right|.
\]

The range \([0 \ldots 1.2]\) is wide enough for most possible yield values to fall into. To this end, a look-up table has been established after extensive numerical simulation; see Table 2. The mean absolute percentage error (MAPE) was less than 5%. The MAPE will not be a serious problem if it is easier to find the global optimal solution after approximation.

For some frequently used values especially, the fitted polynomials are shown below:
\[
x^{1.5} \approx 0.5027x^2 + 0.5308x - 0.0347,
\]
\[
x^{2.5} \approx 1.4357x^2 - 0.4847x + 0.0528,
\]
\[
x^{3} \approx 1.8000x^2 - 0.8953x + 0.1051,
\]
\[
x^{3.5} \approx 2.0997x^2 - 1.2322x + 0.1508,
\]
\[
x^{4} \approx 2.3440x^2 - 1.5053x + 0.1890.
\]

The application method of Table 2 is shown in Figure 2. After approximation, the following QP problems are solved instead.
Take the case in Chen and Wang [9] as an example; see Table 3. In NLP models, the four parameters $o_k, s_k, m_k$, and $d_k$ were set to 1.1, 0.4, 1, and 0.71 in their study. The NLP models were converted into the QP ones. Then both methods were applied to predict the yield. After defuzzifying the yield forecasts with the centroid-of-gravity formula, the forecasting performances by the two methods were compared in Table 4.

(1) Obviously, the forecasting accuracy of the QP models was better. Considering MAPE, the proposed QP-based method reduced the prediction error up to 10%.

(2) On the other hand, it is also possible to use the proposed QP-based method to improve the prediction precision measured in terms of the average range of forecasts.

(3) In theory, as long as quadratic polynomials can very precisely fit the original nonlinear objective functions and constraints, the prediction accuracy

![Table 2: The look-up table.](image)

<table>
<thead>
<tr>
<th>$h$</th>
<th>$\alpha$</th>
<th>$\beta$</th>
<th>$\gamma$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.00</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>1.01</td>
<td>0.0093</td>
<td>0.9927</td>
<td>-0.0020</td>
</tr>
<tr>
<td>1.02</td>
<td>0.0187</td>
<td>0.9852</td>
<td>-0.0039</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
<tr>
<td>4.00</td>
<td>2.3440</td>
<td>-1.5053</td>
<td>0.1890</td>
</tr>
</tbody>
</table>

QP problem I is

$$\text{Min } Z_3 = \sum_{t=1}^{T} \left( \alpha(o(k)) \left( \ln Y_{t3} - \ln Y_{t1} \right)^2 + \beta(o(k)) \left( \ln Y_{t3} - \ln Y_{t1} \right) + \gamma(o(k)) \right),$$

subject to

$$\ln y_t \geq \ln Y_{t1} + s_t \left( \ln Y_{t2} - \ln Y_{t1} \right),$$
$$\ln y_t \leq \ln Y_{t3} + s_t \left( \ln Y_{t2} - \ln Y_{t3} \right),$$

$$\ln Y_{t1} = a_1 - \frac{b_3}{t},$$
$$\ln Y_{t2} = a_2 - \frac{b_2}{t},$$
$$\ln Y_{t3} = a_3 - \frac{b_1}{t},$$
$$0 \leq a_1 \leq a_2 \leq a_3,$$
$$0 \leq b_1 \leq b_2 \leq b_3,$$
$$t = 1 \sim T.$$ (14)

QP problem II is

$$\text{Max } Z_4 = \sum_{t=1}^{T} \left( \alpha(m(k)) s_t^2 + \beta(m(k)) s_t + \gamma(m(k)) \right)$$

subject to

$$\sum_{t=1}^{T} \left( \alpha(o(k)) \left( \ln Y_{t3} - \ln Y_{t1} \right)^2 + \beta(o(k)) \left( \ln Y_{t3} - \ln Y_{t1} \right) + \gamma(o(k)) \right) \leq T \cdot d_k a_k,$$

$$\ln y_t \geq \ln Y_{t1} + s_t \left( \ln Y_{t2} - \ln Y_{t1} \right),$$
$$\ln y_t \leq \ln Y_{t3} + s_t \left( \ln Y_{t2} - \ln Y_{t3} \right),$$

$$\ln Y_{t1} = a_1 - \frac{b_3}{t},$$
$$\ln Y_{t2} = a_2 - \frac{b_2}{t},$$
$$\ln Y_{t3} = a_3 - \frac{b_1}{t},$$
$$0 \leq a_1 \leq a_2 \leq a_3,$$
$$0 \leq b_1 \leq b_2 \leq b_3,$$
$$t = 1 \sim T.$$ (16)
of the proposed QP-based method must be better than the original NLP-based method. Even with little error, a similar effect still exists. For this reason, it seems that it is not necessary to compare the subsequent collaboration step. Fuzzy collaborative forecasting methods based on improved individual forecasts should perform better.

(4) It is also easy to prove that the quadratic objective function and constraints are convex, because \( \ln Y_{t+1} \geq \ln Y_t \) and \( a(o(k)) \geq 0 \). That is very important to the global optimality of the optimal solution.

### 4. Discussion and Conclusion

Fuzzy collaborative forecasting methods have considerable potential for the semiconductor yield forecasting. In this field, the existing methods are mostly based on NLP models. Such a practice has some drawbacks. To solve these problems, we proposed a QP-based fuzzy collaborative forecasting method and made a fair comparison with instances from the literature.

According to the experimental results, the QP-based approach achieved a better prediction performance than the NLP-based method, because it is easier to find the global optimal solution in the QP-based method than the method based on NLP. In addition, the disagreements between the views of different experts are also better distinguished in the QP-based approach. Furthermore, the NLP models are in fact formulated subjectively and certainly can be replaced by any other formulation, to achieve a better performance. The error of estimating the NLP model with the QP models becomes unimportant as well. In other words, the future focus should be on the development of better QP models, rather than the approximation of the NLP models.

### Acknowledgment

This work was supported by the National Science Council of Taiwan.

### Table 3: The case in Chen and Wang [9].

<table>
<thead>
<tr>
<th>( t )</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
</tr>
</thead>
<tbody>
<tr>
<td>( Y_t )</td>
<td>37.30%</td>
<td>58.50%</td>
<td>54.10%</td>
<td>74.10%</td>
<td>61.70%</td>
<td>80.00%</td>
<td>71.20%</td>
</tr>
</tbody>
</table>

### Table 4: The performances by the two methods.

<table>
<thead>
<tr>
<th>Model</th>
<th>The average range (precision)</th>
<th>Mean absolute percentage error (accuracy)</th>
</tr>
</thead>
<tbody>
<tr>
<td>NLP I</td>
<td>0.34</td>
<td>14.5%</td>
</tr>
<tr>
<td>NLP II</td>
<td>0.37</td>
<td>17.2%</td>
</tr>
<tr>
<td>QP I</td>
<td>0.40</td>
<td>14.1%</td>
</tr>
<tr>
<td>QP II</td>
<td>0.28</td>
<td>14.3%</td>
</tr>
</tbody>
</table>

### References


