Statistical modeling in manufacturing: Adapting a diagnostic tool to real-time applications

by B. E. Osborn

This paper describes a process for constructing a statistical model to automate the analysis of data from complex diagnostic tools. The method is demonstrated on data taken from an optical emission spectrometer (OES), one of the most powerful tools used in semiconductor manufacturing for detecting the chemical composition and impurity levels in plasma processes. The analysis of OES data currently requires hours of manual effort by an expert spectroscopist, rendering it ineffective for real-time monitoring and control. However, through the use of statistical modeling, the analysis can be performed automatically on a personal computer in a matter of seconds. The process of model construction is examined in general, and methods are developed for demonstrating how information from an expert can be combined with information from the data in order to provide a statistical basis for analysis. The effectiveness of the model is demonstrated on data from typical plasma processes.

Introduction

• *Background*

One of the most effective ways of influencing the quality of our products is to ensure consistency and reproducibility in the manufacturing environment. This is especially true in semiconductor manufacturing, where small variations in the plasma processes for etching and deposition can have drastic effects on the quality of what is produced.

Statistical process control (SPC) [1] is a valuable tool for accomplishing this objective by comparing present performance with the past and by differentiating between normal statistical variation and process alteration. Unfortunately, many of the sophisticated tools used to diagnose problems in the manufacturing environment do

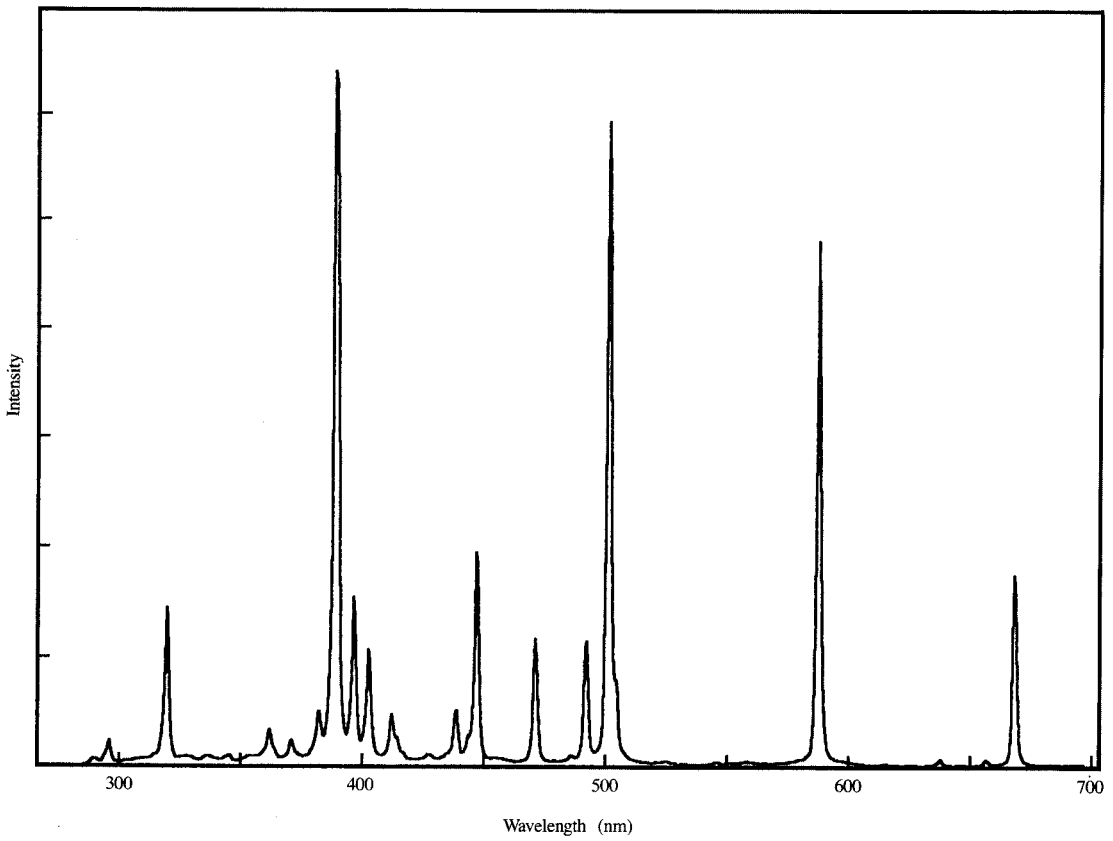

Figure 1

OES spectrum of gas from a high-purity helium bottle.

Table 1 Sample OES peaks for helium.

Wavelength (nm)	*Relative intensity*
388.865	500
396.4729	20
402.6191	50
447.1479	200
501.5678	100
587.562	500
587.597	100
667.815	100

not lend themselves to classical SPC techniques. For example, optical emission spectroscopy (OES), residual gas analysis (RGA), Fourier transform infrared spectroscopy (FTIR), and laser-induced fluorescence (LIF) are all examples of analytical tools which have strong potential for use as monitors in plasma processes, each providing important information about the chemical environment. However, the analysis of data produced by these tools requires expertise and can be a time-consuming process that limits the tool's value for real-time process control applications.

This paper describes a methodology for automating this data analysis process. The focus is on OES because of its relative simplicity and its usefulness in detecting process and tool contamination during semiconductor manufacturing (see [2–4] for a description of how OES works and its applicability to plasma-processing diagnostics). The statistical techniques that are introduced are generally applicable and can be used to analyze data from other spectroscopic tools.

- *Description of the problem*

OES monitors light emitted in the visible region of the spectrum from electronic transitions of atoms and

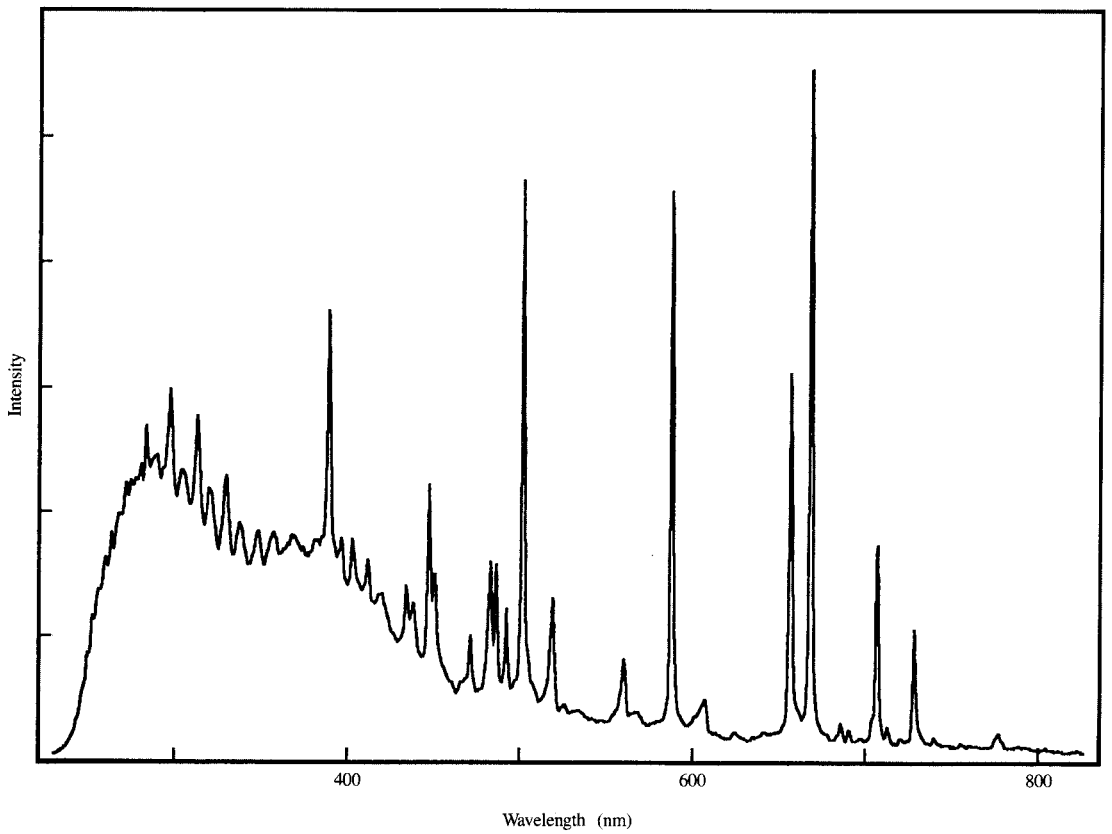

Figure 2

OES spectrum taken during a reactive ion etching process involving a CF_4/CHF_3/He plasma.

molecules with a series of calibrated photodiodes, each measuring the intensity at a specific wavelength. A spectrum taken from a helium gas bottle obtained with a calibrated OES spectrometer is displayed in **Figure 1**. The data are obtained from a *low-resolution* OES spectrometer in which the wavelengths assigned to diodes differ by intervals of about 0.6 nm. Higher-resolution OES spectrometers require longer processing times which limit their practical use for many real-time applications. Although the analysis described below is also applicable to interpretation of high-resolution OES spectra, we focus on analysis of low-resolution OES data.

A particular gas is identified from its OES spectrum by correlating wavelengths at which peaks appear with previously reported wavelengths tabulated for the gaseous species. Part of the tabulation for spectral lines of helium gas (from [5]) is shown in **Table 1**. The relative peak intensities can vary in a particular spectrum, depending on factors which include the power level, the relative concentration of the excited species, the presence of other gases, and variations in the optics. For example, in Figure 1 the peak intensity at 501.5678 is greater than that at 667.815, but both have the same relative intensity in Table 1.

This paper discusses a procedure to build a system that automates the work of an expert OES spectroscopist so that the gas peaks in OES spectra can be properly identified.

- *Previous work*

One of the simplest methods of automating OES data interpretation is by direct comparison with tabulated data. A "table-lookup" procedure is used to identify the species present in an OES spectrum obtained from a RIE tool which utilizes a CF_4/CHF_3/He plasma during semiconductor manufacturing (see **Figure 2**). A gas is

assigned to a particular peak if it is located within 0.6 nm of a corresponding table entry. The table values for 23 gases were used in this process, with 5–79 table entries associated with each gas. The results are not satisfactory. The 0.6-nm tolerance used in this algorithm causes overlap between the tabulated wavelength values and, as a result, *every* gas is found at least once.

Improvements in this technique were made using a specially designed library of waveform peaks corresponding to the individual chemical species [6]. This method takes into consideration the multiplicity of peaks associated with each species and their specific shape in addition to wavelength location. A cross-correlation function is used to determine which library entries are the best matches for peaks in the spectra under investigation. However, variations may arise when the technique is implemented on different spectrometers under various process conditions. In addition, this method fails to consider the user's prior knowledge concerning the presence of various species.

Researchers are currently investigating more sophisticated models employing expert systems [7] and neural networks [8] for species identification. An expert system approach involves coding a set of rules which attempt to duplicate the expertise used by the spectroscopist. An artificial neural network model "learns" to interpret OES data after being exposed to a large number of properly interpreted spectra. Each technique models one important characteristic of the problem: in the case of an expert system, the prior knowledge of the expert, and in the case of the neural network, the need to learn with experience. However, neither approach adequately incorporates both.

• *Designing an appropriate model*

An ideal model for solving the problem of interpreting OES data should incorporate

1. Any prior knowledge concerning which gases are most likely to be present.
2. The expert's knowledge of the problem.
3. The inherent uncertainty in the problem leading to the model's ability to indicate the likelihood of the results.
4. The knowledge that has been gained through experience.

This paper introduces a new statistical methodology for approaching this problem which effectively meets these model criteria. It begins with a simple Bayesian formulation of the problem similar to that used in other pattern-recognition problems [9], adding complexity to construct the desired model.

Components of the statistical model

• *Bayesian pattern recognizer*

Consider a mixture which may contain any combination of N gaseous species. Let $G_i = 1$ indicate that the ith gaseous species is present and let $G_i = 0$ indicate that it is absent. Also, let $p_j = 1$ or 0 be used to indicate whether the jth table entry for this gas is present or absent in the graph ($j = 1, \cdots, m$). First assume that we are dealing with a high-resolution OES spectrum in which peaks correspond *exactly* to the appropriate table entry values. We later relax this assumption to consider an actual low-resolution OES spectrum. From the Bayesian theorem,

$$P[G_i|p_1, p_2, \cdots, p_m] = \frac{1}{Z} P[p_1, p_2, \cdots, p_m|G_i] \times P[G_i], \quad (1)$$

where $P[\cdot|\cdot]$ indicates conditional probability; $P[G_i]$ is the *a priori* probability that the gas is present or absent; and Z is the normalization constant. The conditional probability that certain peaks will be observed in the graph given that the gas is present should depend on the tabulated relative intensities. It is much more likely that we will observe a peak with a relative intensity value of 500 than one with a relative intensity of 10. An appropriate assumption is that

$$P[p_1, p_2, \cdots, p_m|G_i] = P[p_1|G_i] \times P[p_2|G_i] \times \cdots \times P[p_m|G_i], \quad (2)$$

where each $P[p_j|G_i = 1]$ is some function of the relative intensity for table entry j. The conditional probability that a particular peak would be detected even if the corresponding gas were not present (i.e., $P[p_j = 1|G_i = 0]$) is a small constant based on noise in the system. The specific value of this constant can be determined by using maximum-likelihood or similar parameter-estimation techniques on a set of *learning* data in a system in which the gases present are known.

Once the parameters have been estimated and there exists an unknown gas mixture, the Bayesian classifier determines the presence of the gas in question if

$$P[G_i = 1|p_1, p_2, \cdots, p_m] > P[G_i = 0|p_1, p_2, \cdots, p_m]. \quad (3a)$$

Combining Equations (1) and (2) results in the equivalent expression:

$$P[p_1|G_i = 1] \times P[p_2|G_i = 1] \times \cdots \times P[p_m|G_i = 1] \times P[G_i = 1] > P[p_1|G_i = 0] \times P[p_2|G_i = 0] \times \cdots \times P[p_m|G_i = 0] \times P[G_i = 0]. \quad (3b)$$

• *Pattern theory model*

The above model would be appropriate if there existed a perfect high-resolution OES where each peak could be observed at its exact wavelength. In real systems, this is

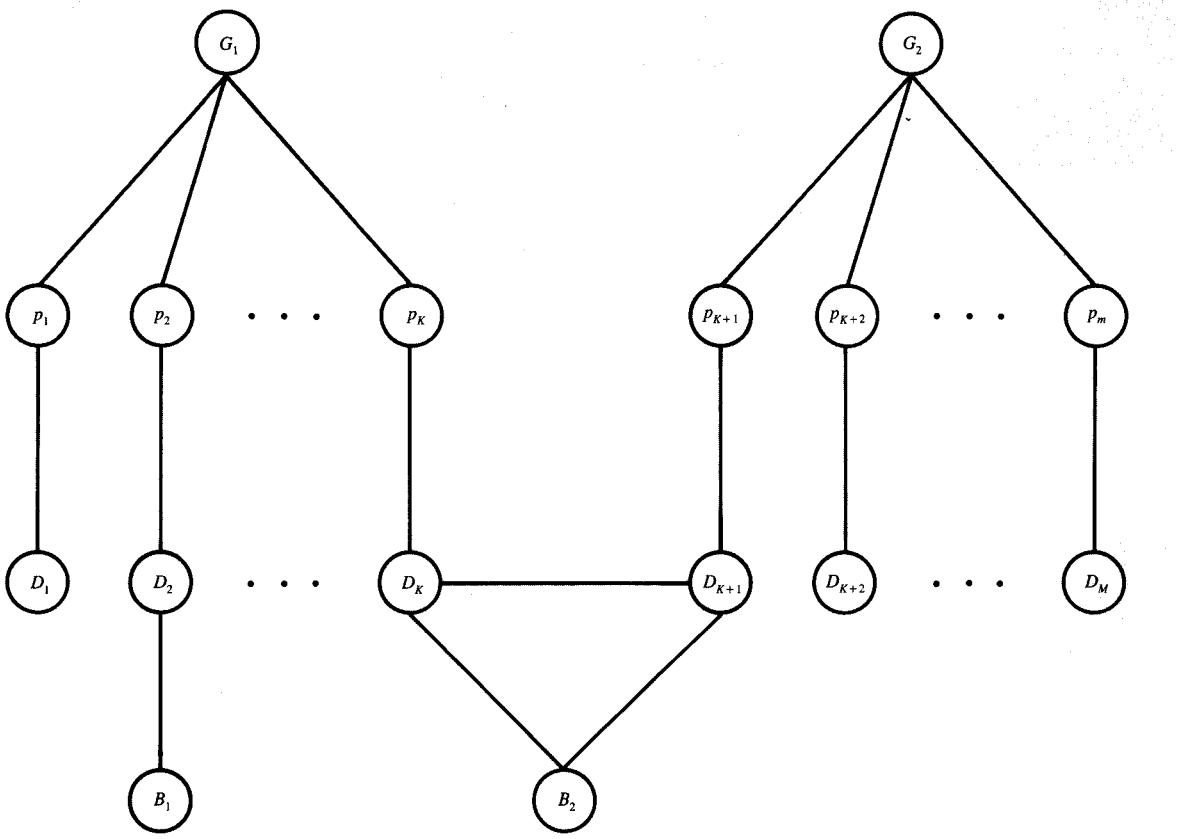

Figure 3

Graphical representation of the general pattern model used to analyze OES data.

not the case. If the peak associated with the jth table entry were present in the graph, it could be observed at a wavelength different from the one tabulated. This difference would depend on the level of resolution of the OES spectrometer and the accuracy of the instrument calibration procedure. In addition, other gases in the system affect which peaks of the gas under study are detected. It is extremely difficult to consider all of the complexities involved, since these effects relate to the basic physics of the OES process. However, a more realistic and accurate model can be constructed.

From a statistical point of view, the issues raised in the preceding paragraph can be summarized by saying that the values $p_1, p_2, \cdots, p_m$ cannot be observed directly. Instead, the peaks in the graph represent some *deformed image* of these values. For simplicity, consider two types of *deformation mechanisms* that alter the p_j's:

1. A simple shift in the wavelength. For example, if $p_j = 1$, the peak is not observed at its true wavelength, but rather at another wavelength.
2. A *blurring* effect, in which two distinct peaks at wavelengths λ_1 and λ_2 are observed as a single peak at wavelength α.

A model which takes all of the above [including Equations (1)–(3)] into consideration is represented graphically in **Figure 3**. This is an example of a *pattern theory model* whose general structure is detailed by Grenander [10, 11].

The circles in Figure 3 are referred to as *sites*, and the line connecting the circles are referred to as *segments*. There are four levels of sites displayed. At the top level are the G sites associated with the gases that may exist in the mixture under study. (For simplicity, only two G sites are shown.) The sites at the second level are the p sites

associated with the table values for these gases. The sites at the third level are the D sites and are the result of the first distortion mechanism. Finally, the B sites at the fourth level are the result of the second distortion mechanism and are associated with the observed peaks in the graph.

The G sites take values of 1 or 0 depending on whether the associated gas is present or absent. Similarly, each of the p sites takes a value of 1 or 0 depending on whether or not the peak associated with the particular table entry is present or absent in the graph. The D sites at the third level and the B sites at the fourth level all take values from the positive real line.

A site D_j connects to a site B if $|\omega_j - B| < L$, where ω_j is the wavelength of the table entry associated with p_j and L is a constant. We note that if L is less than the level of resolution of the graph (i.e., the distance between adjacent diodes), each D site connects to, at most, one B site, since at least three diodes are required to define a peak.

When a B site connects to more than one D site, the B site represents a "blurred" wavelength that results from distortion mechanism (2). Note that this blurring deformation occurs only if wavelengths of corresponding peaks are sufficiently close. In the case where a B site connects to only a single D site, both of the sites have the same value, and no blurring distortion occurs. Finally, if a D site does not connect to a B site, the value of the D site is fixed at 0 [hence, $D_j \in \{0, (\omega_j - L, \omega_j + L)\}$]. In the example displayed in Figure 2, the Kth table entry associated with the first gas is close to the wavelength for the $K + 1$th table entry associated with the second gas.

Therefore, the problem of interpreting OES data (e.g., Figure 2) using this model reduces to a statistical problem of determining the G, p, and hidden D sites given the values of the B sites.

As one might have concluded from the above formulation, the segments in the graph represent *relationships* between the connecting sites. More precisely, the segments indicate the *conditional probabilities inherent in the pattern theory model.* Specifically, the conditional probability of the value of any site, given the values of all other sites in the graph, is equal to the conditional probability of the value of the site, given the values of just those sites which are directly connected to the subject site. For example, considering the site p_1, which is connected to the sites G_1 and D_1,

$$P[p_1|\text{all other sites in the graph}] = P[p_1|G_1, D_1]. \tag{4}$$

This is called the *Markovian* relationship, and mathematical structures which possess such a relationship are *Markov random fields* [12].

Model construction

The general probability measure for pattern models is described in [11]. In our case, the probability density function for the general model having N gases and $(M_i - M_{i-1})$ table entries for gas i, and where the number of B sites is designated by S, is given by

$$f[G_1, \cdots, G_N, p_1, \cdots, p_{M_N}, D_1, \cdots, D_{M_N}, B_1, \cdots, B_S]$$
$$= \frac{1}{Z} \prod_{i=1}^{N} \prod_{j=M_{i-1}+1}^{M_i} A_{ji}(p_j, G_i) \times H(p_j, D_j)$$
$$\times \prod_{t=1}^{S} Q_t(D_{t_1}, \cdots, D_{t_{k(t)}}, B_t), \tag{5a}$$

where sites $D_{t_1}, \cdots, D_{t_{k(t)}}$ connect to site B_t, $M_0 = 0$, and Z is the appropriate normalizing constant.

If there were no multiply connected B sites, Equation (5a) would simplify to

$$f[G_1, \cdots, G_N, p_1, \cdots, p_{M_N}, D_1, \cdots, D_{M_N}, B_1, \cdots, B_S]$$
$$= \begin{cases} \bar{f}[G_1, \cdots, G_N, p_1, \cdots, p_{M_N}, D_1, \cdots, D_{M_N}] \\ \qquad\qquad \text{for } B_t = D_t,\ t = 1, \cdots, S, \\ 0 \qquad\qquad \text{otherwise,} \end{cases}$$

where

$$\bar{f}[G_1, \cdots, G_N, p_1, \cdots, p_{M_N}, D_1, \cdots, D_{M_N}]$$
$$= \frac{1}{\bar{Z}} \prod_{i=1}^{N} \prod_{j=M_{i-1}+1}^{M_i} A_{ji}(p_j, G_i) \times H(p_j, D_j) \tag{5b}$$

and $\bar{Z}$ is the appropriate normalizing constant.

The functions $A(\cdot, \cdot)$, $H(\cdot, \cdot)$, and $Q_t(\cdot, \cdots, \cdot)$ are called *acceptor functions*, and the G, p, D, and B variables are referred to as *generators* which take values from some set of *generator spaces*. In this case, $G_1, \cdots, G_N$ and $p_1, \cdots, p_{M_N}$ take values from discrete generator spaces, and $D_1, \cdots, D_{M_N}$ and $B_1, \cdots, B_S$ take values from continuous generator spaces. Generators are associated with the sites in the graph, and acceptor functions are associated with the segments in the graph.

We are dealing with a pattern model of a partially homogeneous graph [11], since it contains unique acceptor functions [$A_{ji}(\cdot, \cdot)$ and $Q_t(\cdot, \cdots, \cdot)$] and a repeated acceptor function $H(\cdot, \cdot)$. In the next section we discuss the restrictions necessary for identifying the acceptor functions. However, two restrictions can be introduced here:

1. The parameters in Equation (5) must be specified in such a way that the probability is positive for all possible configurations, i.e., all possible values of $(G_1, \cdots, G_N, p_1, \cdots, p_{M_N}, D_1, \cdots, D_{M_N}, B_1, \cdots, B_S)$. This is referred to as the *positivity condition* [13].

2. We wish to construct our model in such a way as to favor configurations for which $D_j \neq 0$ when $p_j \neq 0$ and disfavor configurations for which $D_j \neq 0$ when $p_j = 0$. It is therefore required that

$$H(1, 0) < H(1, D) \text{ and } H(0, D) < H(0, 0)$$
$$\forall\, D \in (\omega_j - L, \omega_j + L).$$

If we define $\mathbb{P}$ as a particular configuration of the B sites and $\mathbb{C}$ as a particular configuration of the G, p, and D sites, the statistical problem is to find the value of $\mathbb{C}$ which maximizes

$$f[\mathbb{C}|\mathbb{P}] \equiv f[G_1, \cdots, G_N, p_1, \cdots, p_{M_N}, D_1, \cdots, D_{M_N}|B_1, \cdots, B_S]. \quad (6)$$

This maximizing value is referred to as $\mathbb{C}_{\max}(\mathbb{P})$.

The classical method for finding $\mathbb{C}_{\max}(\mathbb{P})$ is through *stochastic relaxation* [14]. Basically, this method involves visiting each of the hidden sites in turn and updating their values by choosing a random number from the respective conditional probability distribution. As this process is continued, a distribution of configurations is generated in which the most likely configurations appear most frequently. Stochastic relaxation is a computationally intensive procedure and, if used on this problem, would require several hours of computing time on a personal computer. To overcome this difficulty, we introduce an alternative method which proves to be more efficient when certain restrictions are introduced. First, however, we discuss the problem of parameter identification.

Model implementation

• *Parameter identification*

In order to implement the model, the acceptor functions in Equation (5) must be identified. We use general information concerning the parameters in pattern theory models [11] and knowledge about the structure and physics of the OES system ("expert" knowledge) to assist us in this procedure.

We begin with the acceptor function $Q_t(\cdot, \cdots, \cdot)$. We introduce a simple restriction to simplify our problem.

Theorem 1 For each B_t, let Λ_t represent some deterministic function of the values of the $k(t)$ connecting D sites, and let

$$\Delta(B_1, \cdots, B_S) = \{\mathbb{C}: B_t = \Lambda_t \ \forall\ t\}.$$

Furthermore, define $\tilde{\mathbb{C}}_{\max}(\Delta)$ as the value of $\mathbb{C}$ which maximizes Equation (5b) subject to $\mathbb{C} \in \Delta(B_1, \cdots, B_S)$. Then, if Q_t is defined as

$$Q_t(D_{t_1}, \cdots, D_{t_{k(t)}}, B_t) = \begin{cases} q_t(B_t) & \text{if } B_t = \Lambda_t, \\ r_t(D_{t_1}, \cdots, D_{t_{k(t)}}, B_t) & \text{otherwise,} \end{cases} \quad (7)$$

such that

$$\frac{q_t(B_t)}{r_t(D_{t_1}, \cdots, D_{t_{k(t)}}, B_t)} > \Gamma^{k(t)}, \quad (8)$$

where

$$\Gamma = \max\left[\frac{H(0, 0)}{H(0, x)}, \frac{H(1, x)}{H(1, 0)}\right] \quad \text{maximized for } x \in (\omega_j - L, \omega_j + L),$$

then

$$\mathbb{C}_{\max}(\mathbb{P}) = \tilde{\mathbb{C}}_{\max}(\Delta).$$

The proof of Theorem 1 is in the Appendix.

Theorem 1 tells us that under certain conditions the problem of finding the configuration which maximizes Equation (5a) reduces to the much simpler problem of finding the configuration which maximizes Equation (5b). As a typical implementation of the model, we select the value of Q so that B_t represents a *weighted average* of connecting D sites. For example, consider

$$\Lambda_t = \frac{\sum_{j=t_1}^{t_{k(t)}} D_j \times I_j \times \{D_j \neq 0\}}{\sum_{j=t_1}^{t_{k(t)}} I_j \times \{D_j \neq 0\}}, \quad (9)$$

where I_j is the intensity value for the table entry associated with the jth peak, and where $\{\cdot\}$ is the indicator function which equals 1 if the enclosed expression is true and equals 0 otherwise.

Next consider the acceptor function $H(\cdot, \cdot)$. When $p_j = 1$, D_j represents a shift in the observed value of the wavelength due to noise that is present in the system. This is observed during the calibration process, when the spectrum for a known gas is produced and the peaks are associated with known tabulated data by using a regression procedure. Since the resulting residual error represents the sum of many factors, the central limit theorem suggests that the error distribution is normal. Therefore,

$$f[D_j = x|p_j = 1] = \begin{cases} \left(\frac{1}{\sqrt{2\pi\sigma}} \times \exp\left[-\frac{(x - \omega_j)^2}{2\sigma^2}\right]\right) & \text{for } x \in (\omega_j - L, \omega_j + L), \\ K & \text{for } x = 0, \end{cases} \quad (10)$$

where K is the normalizing constant. Without loss of generality, we can choose $K = e^b/(1 + e^b)$; K therefore

equals the probability that we do not observe the peak in the graph even though it should be present (i.e., $p_j = 1$). This may be due to a number of reasons, the most likely of which is that it is outside our region of interest, i.e., outside $(\omega_j - L, \omega_j + L)$. For example, if we choose L to be 3σ, the probability that D_j will fall outside the region is 0.0026, and therefore $b = -5.95$. In addition, experimental results indicate that an appropriate value for σ is 0.2.

In the case where $p_j = 0$, D_j would take on nonzero values because of statistical uncertainties such as noise in the system and stray peaks from gases not being considered. It is reasonable to assume that these nonzero values of D_j would be uniformly distributed over the region of interest; hence, we can write this conditional probability density function as

$$f[D_j = x|p_j = 0] = \begin{cases} \left[\dfrac{e^a}{2 \times L \times (1 + e^a)}\right] & \text{for } x \in (\omega_j - L, \omega_j + L), \\ \left(\dfrac{1}{1 + e^a}\right) & \text{for } x = 0. \end{cases} \quad (11)$$

For a noise level of 10% we would choose $a = -2.197$.

With this information and the fact that conditions on identifiability restrict $H(0, 0) = 1$ [11], we find that Equations (10) and (11) imply

$$\begin{aligned} &H(0, 0) = 1, \\ &H(1, 0) = e^b, \\ &H(0, x) = \left(\frac{e^a}{2 \times L}\right) \quad \text{for } x \neq 0, \\ &H(1, x) = \left(\frac{1 + e^b}{\sqrt{2\pi}\sigma} \times \exp\left[-\frac{(x - \omega_j)^2}{2 \times \sigma^2}\right]\right) \quad \text{for } x \neq 0. \end{aligned} \quad (12)$$

Two restrictions are necessary on $A_{ji}(\cdot, \cdot)$ [11]:

$A_{ji}(0, 0) = 1$ and $A_{ji}(0, 1) = 1$ for $j \neq 1$.

Without loss of generality we can define

$$\begin{aligned} &A_{ji}(0, 0) = 1, \\ &A_{ji}(1, 0) = e^{\varepsilon_j}, \\ &A_{ji}(0, 1) = \begin{cases} e^{\delta_i} & \text{if } j = 1, \\ 1 & \text{otherwise,} \end{cases} \\ &A_{ji}(1, 1) = \begin{cases} e^{\delta_i + \varepsilon_j + f_j} & \text{if } j = 1, \\ e^{\varepsilon_j + f_j} & \text{otherwise.} \end{cases} \end{aligned} \quad (13)$$

The resulting conditional probability is given by

$$P[G_i = 1|p_{M_{i-1}+1}, \cdots, p_{M_i}] = \frac{\exp\left(\delta_i + \sum_{j=M_{i-1}+1}^{M_i} p_j \times f_j\right)}{1 + \exp\left(\delta_i + \sum_{j=M_{i-1}+1}^{M_i} p_j \times f_j\right)}. \quad (14)$$

Here, there is a term f_j associated with each table entry value. This will be some function of the relative intensity value. One that works quite well in practice is

$$f_j = 7.3244 + (0.018847 \times K_j),$$

where

$$K_j = \begin{cases} I_j & \text{if } I_j < 100, \\ 100 & \text{if } 100 \leq I_j < 150, \\ 150 & \text{if } 150 \leq I_j < 250, \\ 225 & \text{if } I_j \geq 250. \end{cases}$$

There is also a term δ_i associated with each gas which, as soon becomes apparent, need not be estimated to solve our problem.

Using Equations (12) and (13), we find that the conditional probability of p_j given D_j is given by

$$P[p_j = 1|G_i] = \frac{\exp[\varepsilon_j + (G_i \times f_j)]}{C + \exp[\varepsilon_j + (G_i \times f_j)]}, \quad (15)$$

where

$$C = \frac{1 + e^a}{1 + e^b}.$$

When $G_i = 0$, we would expect the probability that $p_j = 1$ to be extremely small. Furthermore, we would expect this type of "noise" term to be independent of the particular gas or peak under consideration. This implies that $\varepsilon = \varepsilon_j$ $\forall j$. By choosing $\varepsilon = -6.8$, we ensure that $P[P_j = 1|G_i = 0] = 0.001$.

We conclude this section by stating the following proposition, which provides further information on the values of the D sites.

Theorem 2 Under the conditions of Equations (9) and (10), for each B_t, the values of the connecting D sites in $\bar{\mathbb{C}}_{\max}(\Delta)$ are either zero or have the form

$$D_j = \omega_j + \frac{\lambda \times I_j}{\sum_{i=t_1}^{t_{k(t)}} I_i \times \{D_i \neq 0\}}, \quad (16a)$$

where

$$\lambda = \frac{\left(B_t - \dfrac{\sum I_i \times \omega_i \times \{D_i \neq 0\}}{\sum I_i \times \{D_i \neq 0\}}\right) \times \left(\sum I_i \times \{D_i \neq 0\}\right)^2}{\sum I_i^2 \times \{D_i \neq 0\}} \tag{16b}$$

and all sums are between $i = t_1$ and $i = t_{k(t)}$.

The proof of Theorem 2 is in the Appendix.

• *Model solution*

To solve the model, we introduce a variation on the stochastic relaxation algorithm. We make use of the following identity, which follows directly from the Markov property (4) and from Equation (5):

$$\tilde{f}[G_1, \cdots, G_N, p_1, \cdots, p_{M_N}, D_1, \cdots, D_{M_N}]$$

$$= \prod_{i=1}^{N} \prod_{j=M_{i-1}+1}^{M_i} f[D_j|p_j] \times P[p_j|G_i] \times P[G_i]. \tag{17}$$

We now propose the following algorithm.

Theorem 3 (selective stochastic relaxation algorithm) Assume the conditions of Theorems 1 and 2, and define the following four sets:

$\Omega_i = \{j: (M_{i-1} + 1) \leq j \leq M_i\}$,

$\Theta_i = \{B_t: \Omega_i \cap (t_1, \cdots, t_{k(t)}) \neq \varnothing\}$,

where $D_{t_1}, \cdots, D_{t_{k(t)}}$ are the sites which connect to site B_t, and $\varnothing$ indicates the null set;

$\Pi_i = \{(t_1, \cdots, t_{k(t)}): B_t \in \Theta_i\}$,

$N(i) = \{G_l(l \neq i): \Pi_i \cap \Omega_l \neq \varnothing\}$.

Define $G_{(j)}$ as the G site which connects to site p_j, $\mathbf{G}$ as a particular configuration of the G sites, and $D_j(\mathbf{G})$ and $p_j(\mathbf{G})$ as the sites which maximize $f[D_j|p_j]$ for $G_{(j)} \in \mathbf{G}$. Then, the most frequently occurring configuration will be $\mathbb{C}_{\max}(\mathbb{P})$, when the graph is updated using the following algorithm for determining the values of the G, p, and D sites:

1. Initialize the values of $(D_1, \cdots, D_{M_N})$ by assigning the value of each B site to the connecting D site whose corresponding ω_j is closest to the value of the B site. For example, if sites D_1, D_2, and D_3 are connected to B, and if $|\omega_2 - B| < |\omega_3 - B| < |\omega_1 - B|$, then assign $D_2 = B$ and $D_1 = D_3 = 0$ (in the case of a tie, simply choose one to be "closest").
2. Find the values of the G and p sites which maximize Equation (17).
3. Update each G_i, p_j, and D_j site where $j \in \Pi_i \cup \Omega_i$ as follows: Determine two sets of the p_j and D_j sites $(j \in \Pi_i \cup \Omega_i)$ such that Equation (16) holds and
   a. In the first set, maximize Equation (17) for $G_i = 0$.
   b. In the second set, maximize Equation (17) for $G_i = 1$.

   Select the set with which to update the graph according to whether $G_i = 0$ or $G_i = 1$ when it is chosen as a random number from the conditional distribution:

$$\Phi(G_i|G_1, \cdots, G_{i-1}, G_{i+1}, \cdots, G_N) = \frac{\prod_{j \in \Pi_i \cup \Omega_i} f[D_j(\mathbf{G})|p_j(\mathbf{G})] \times P[p_j(\mathbf{G})|G_{(j)}] \times \prod_{G \in N(i)} P[G]}{\sum_{G_i=0}^{1} \prod_{j \in \Pi_i \cup \Omega_i} f[D_j(\mathbf{G})|p_j(\mathbf{G})] \times P[p_j(\mathbf{G})|G_{(j)}] \times \prod_{G \in N(i)} P[G]} \tag{18}$$

4. Once this procedure is repeated on all of the G sites, record the configuration and return to step 3.

The proof of Theorem 3 is in the Appendix.

This method is called *selective* stochastic relaxation because only G sites are updated. By calculating the relative frequency of each configuration encountered, the *relative* probability that each configuration of the G sites will occur can be estimated. The word "relative" is stressed because the distribution $\Phi(\mathbf{G})$ only deals with a subset of the total possible set of configurations of the G, p, and D sites.

• *Examples*

CF_4/CHF_3/He plasma analysis

This algorithm has been implemented in software, it has been tested extensively using OES data collected under various conditions both in the laboratory and in the manufacturing environment, and a patent application for its implementation has been filed. Twenty-three of the gaseous species most relevant to semiconductor manufacturing processing have been included, with 10–80 table entries associated with each gas. A copyrighted version of the software implemented in C is available from the author.

The model runs sufficiently fast for real-time response in the manufacturing environment. For example, less than five seconds are required on a PC with a 486 microprocessor to perform 1000 iterations of the stochastic relaxation algorithm on the data displayed in Figure 2.

Figure 4 shows a graphical view of the model applied to the data displayed in Figure 2. To make the figure more comprehensible, only 12 of the 23 gases are displayed. The x symbols located within the G sites in Figure 4 designate the most frequently occurring configuration that results from the application of the selective stochastic relaxation algorithm. From the algorithm, five gases were found: He, H, CF_2, CO, and F. This result corresponds to the

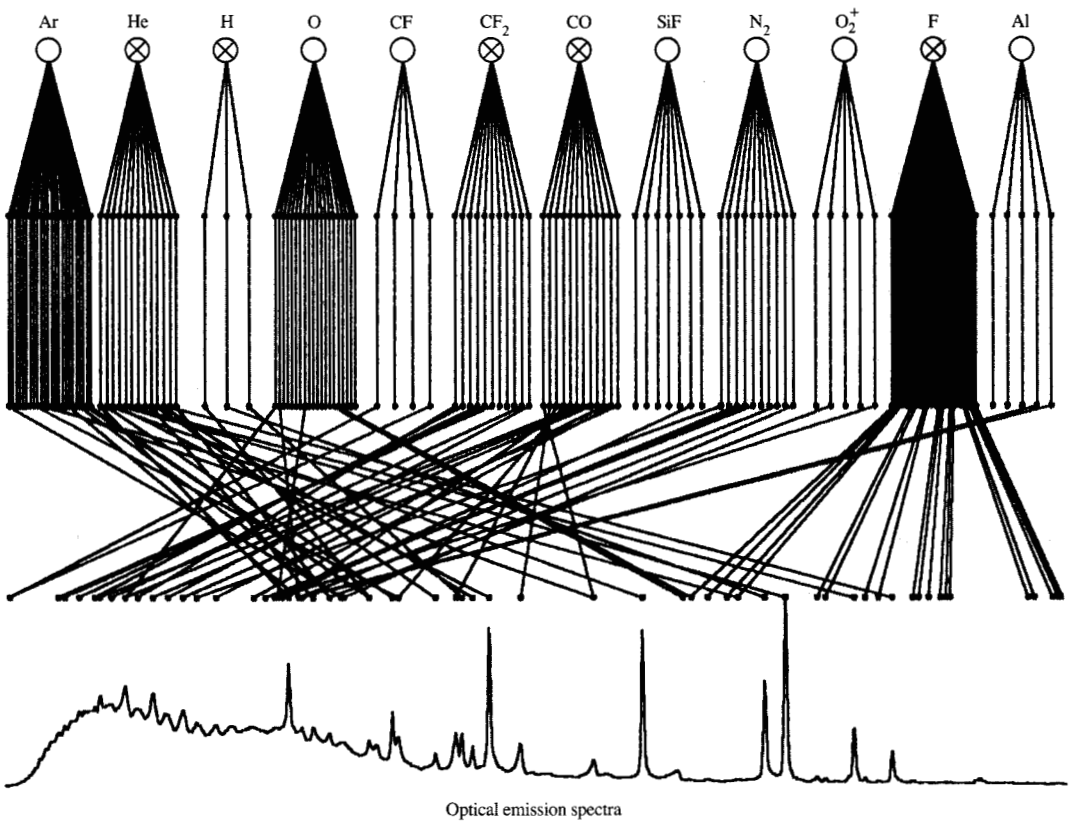

Figure 4

Graphical representation of the pattern model used to analyze the data displayed in Figure 2.

conclusion of an expert spectroscopist. The resulting peak assignments and the relative probabilities of each of the relevant gases are shown in **Figure 5**.

Detecting process impurities

One of the important applications of OES is to detect impurities in the plasma. It is not uncommon to have impurities present during the etching and deposition process, detracting from product quality. At best, these process impurities reduce yield but produce product which can be reworked. At worst, the impurities remain undetected, and the product performs inadequately when integrated into a final system.

Impurities are often introduced through leaks in a system. An air leak is most easily detected by the presence of nitrogen in the OES trace. **Figure 6** is an OES spectrum to test for reproducibility of the plasma shown in Figure 5. Figure 6 shows that the model has clearly detected nitrogen. When installed in a manufacturing system, this model would detect this "out of control" condition and would automatically stop the process to prevent further damage of the product.

Figure 7 illustrates how a nitrogen contaminant was identified in a helium gas bottle guaranteed to be of high purity. This model can be used to qualify gas cylinders before they are connected to manufacturing tools.

Model evaluation

The model meets the model selection criteria defined earlier in the following ways:

1. Prior knowledge concerning which gases are most likely to be present is represented in the *a priori* probabilities $P[G_i]$.
2. The expert's knowledge has been incorporated in the positioning of sites and segments. In addition, actual

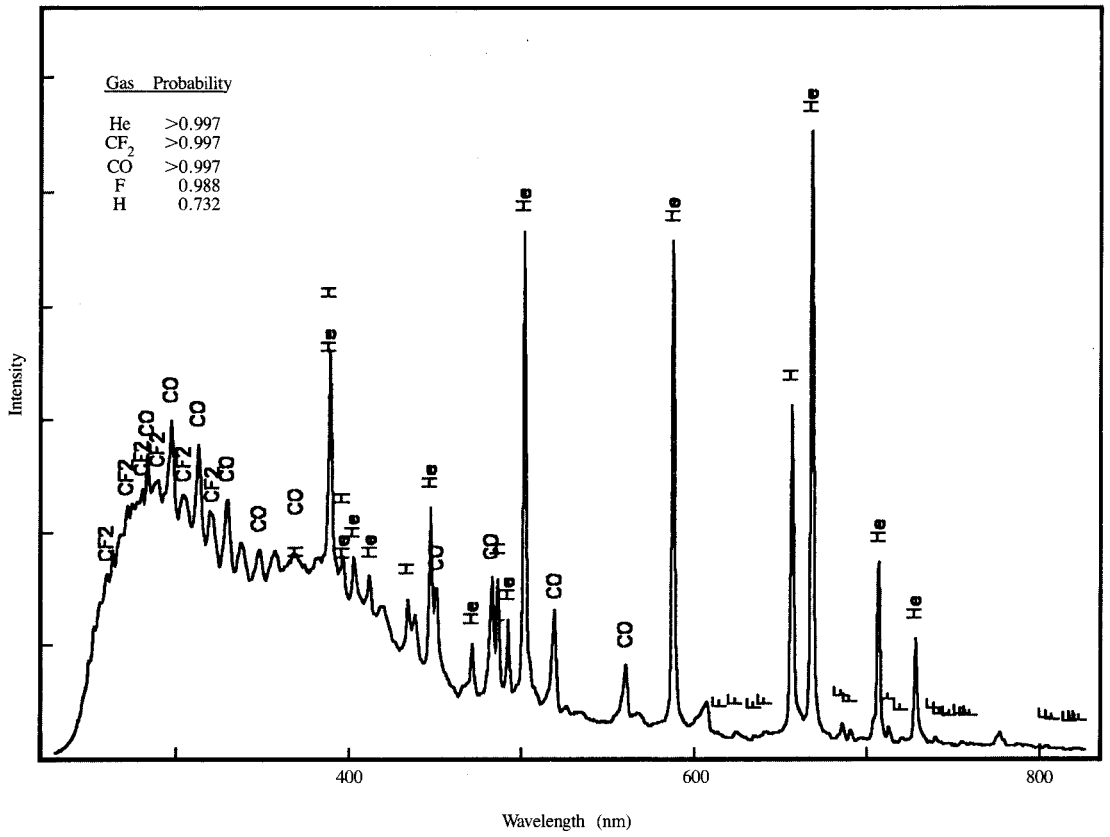

Figure 5

Results of applying the pattern model to the data displayed in Figure 2.

values of the conditional probabilities are derived from an understanding of the calibration procedure and the meaning of the tabulated relative intensity values.

3. The inherent uncertainty in the problem has been taken into account in the stochastic structure of the model. The selective stochastic relaxation algorithm enables the user to determine the relative probability of the possible solutions.
4. Knowledge from experience can be incorporated into the model through updating the conditional probability in Equations (10), (11), and (15).

The model has some shortcomings. First, it is very dependent on which table entries are included. If too few entries for a certain gas are included, recognition of that gas may be difficult in certain instances; if too many entries are included, additional peaks must be detected in order to conclude that a gas is present. As a result, building these tables requires OES expertise and some degree of experimentation. Fortunately, practice has shown the model to be very robust in that gases are recognized fairly easily despite variations in the specification of the parameters and inclusion or exclusion of specific table values. Experiments so far have revealed that the model is more robust for helium-based plasmas than for those with an argon base. This disadvantage is easily overcome through experimentation and through updating the conditional probabilities (i.e., Bayesian learning [9]) as the model is exposed to spectra under various conditions.

A second disadvantage is the fact that fairly precise calibration is necessary before the model can be executed. For the type of low-resolution spectra that are examined, a difference of 0.6 nm can have a drastic effect on the output of the model. One method for solving this problem is through a technique of *dynamic calibration*, in which the model is first run with a larger value of σ, with a specified

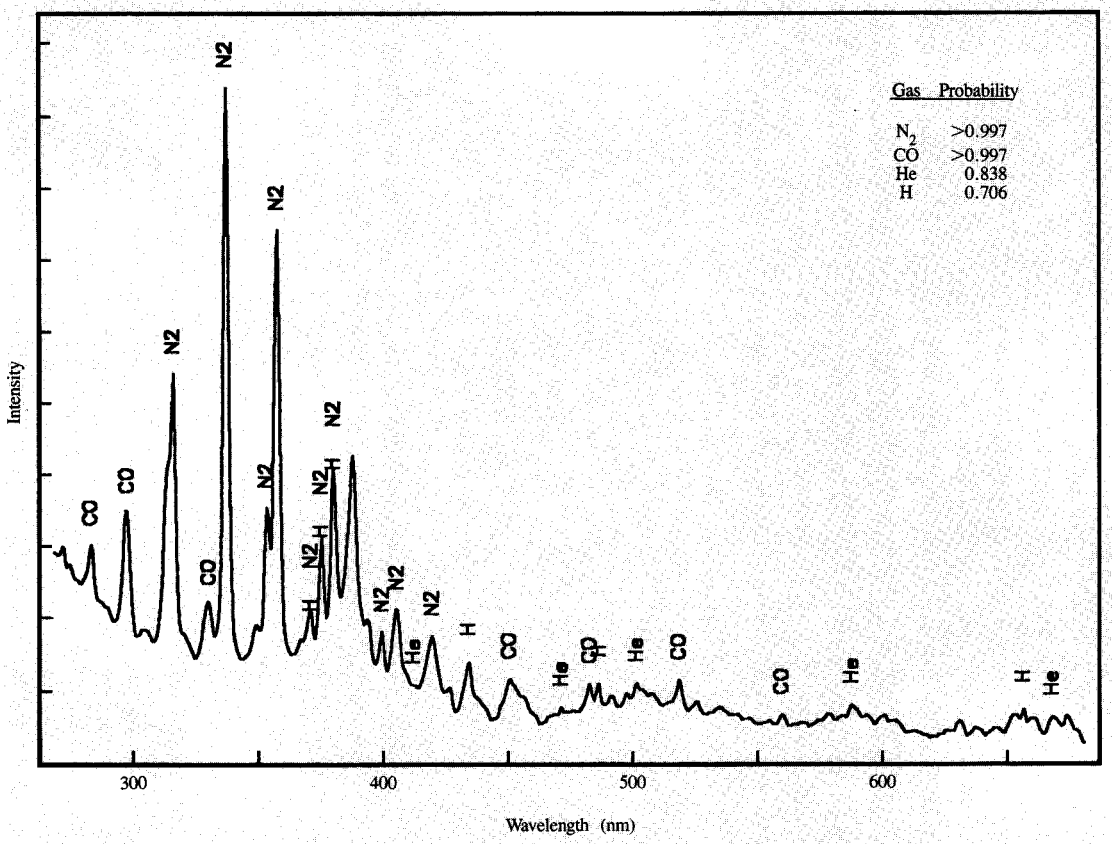

Figure 6

Results of applying the pattern model to OES data taken from a CF_4/CHF_3/He mixture in the laboratory with a nitrogen leak.

gas which is known to be present (e.g., helium for the data shown in Figure 2), and with only a few of the major peaks in the graph considered (e.g., the five most intense peaks). For this run, the prior probability of the specified gas is set to 1.0 and the others are set to zero. Once these first few peaks have been identified, the data are calibrated and the process is continued with all of the peaks (or with a larger subset of the peaks) and with σ restored to its original value (or perhaps some intermediate value). The data are then recalibrated on the basis of the additional peaks identified. Finally, everything is restored to its original value, and the model is run on the now-calibrated data. This method of using successively smaller values of σ and an increasing number of peaks is reminiscent of the method of sieves [15]. This method has been tried for various spectra, and has been shown to be most effective when the calibrating gas produces distinct peaks throughout the spectrum (as helium does in Figure 2). When peaks for the gas exist in only one portion of the spectrum, the resulting calibration may be biased. This technique is currently undergoing further refinement.

Concluding remarks and current work

A statistical model for interpreting optical emission spectroscopy data is described. The method involves combining the expertise of the spectroscopist with knowledge gained from data to determine which gases are present. Because the model is stochastic in nature, probabilities can be assigned to the results, enabling the user to determine the amount of confidence that should be placed in the output. By utilizing standard statistical techniques, the model can use its own output to learn from experience. Examples are also presented from actual manufacturing OES traces and from a laboratory experiment showing the effectiveness of the model in determining gas composition.

The accuracy of the model can be improved by modifying the definition of the p and D sites. In the description above,

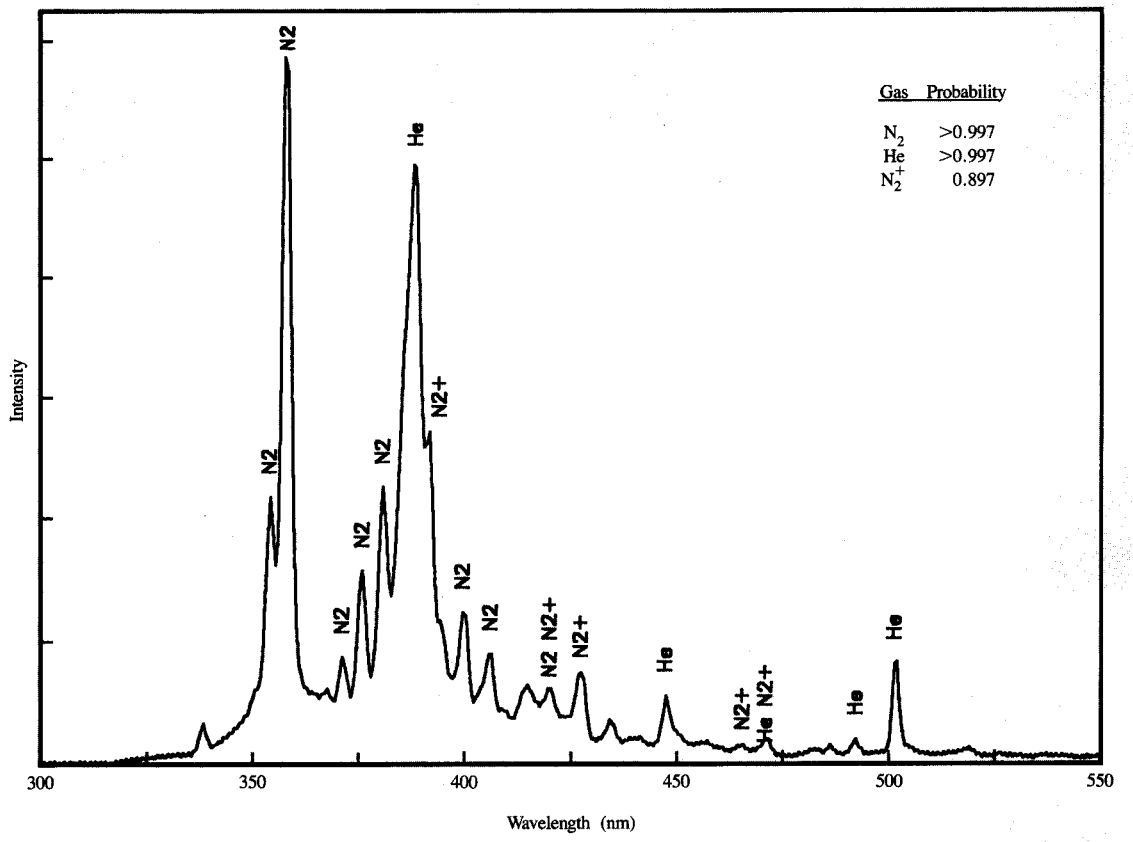

Figure 7

Results of applying the pattern model to OES data taken from a helium bottle which contains a nitrogen contaminant.

the p and D sites contain only information concerning peak locations. Other information can also be contained in these sites, including the relative intensity and the shape of the peak. This latter characterization would help in differentiating between atomic and molecular species.

Finally, the modeling techniques described above are not restricted to the interpretation of OES data. With some minor modifications, this model can be used to analyze data from other complex diagnostic tools in order to ensure quality through real-time monitoring and control in the manufacturing environment. Such work is in progress.

Appendix

Proof of Theorem 1 To prove Theorem 1, the following auxiliary proposition is required.

Lemma 1 If conditions (7) and (8) hold, then $\mathbb{C}_{\max}(\mathbb{P}) \in \Delta(B_1, \cdots, B_S)$.

Proof of Lemma 1 For any $\mathbb{C} \in \Delta(B_1, \cdots, B_S)$,

$$\prod_{i=1}^{N} \prod_{j=M_{i-1}+1}^{M_i} A_{ji}(p_j, G_i) \times H(p_j, D_j) \times \prod_{t=1}^{S} q_t(B_t)$$

$$> \prod_{i=1}^{N} \prod_{j=M_{i-1}+1}^{M_i} A_{ji}(p_j, G_i) \times H(p_j, D_j)$$

$$\times \prod_{t=1}^{S} \Gamma^{K(t)} r_t(D_{t_1}, \cdots, D_{t_{k(t)}}, B_t)$$

$$= \prod_{i=1}^{N} \prod_{j=M_{i-1}+1}^{M_i} \{A_{ji}(p_j, G_i) \times H(p_j, D_j) \times \Gamma\}$$

$$\times \prod_{t=1}^{S} r_t(D_{t_1}, \cdots, D_{t_{k(t)}}, B_t).$$

For $p_j = 0$,

$$H(0, D_j) \times \Gamma \geq H(0, D_j) \times \frac{H(0, 0)}{H(0, D_j)} = H(0, 0).$$

Similarly, for $p_j = 1$,

$$H(1, D_j) \times \Gamma \geq H(1, D_j) \times \max\left[\frac{H(1, x)}{H(1, 0)}\right]$$

$$\geq H(1, D_j) \times \frac{1}{H(1, D_j)} \times \max\left[H(1, x)\right]$$

$$= \max\left[H(1, x)\right],$$

where $x \in (\omega_j - L,\ \omega_j + L)$. Therefore,

$$H(p_j, D_j) \times \Gamma \geq H(p_j, D)\ \ \forall\ D \in \{0, (\omega_j - L, \omega_j + L)\},$$

which implies, for any set of D sites $(D_1, \cdots, D_{M_N})$ such that the configuration

$$(G_1, \cdots, G_N, p_1, \cdots, p_{M_N}, D_1, \cdots, D_{M_N}) \in \Delta(B_1, \cdots, B_S)$$

and any set of D sites $(d_1, \cdots, d_{M_N})$ such that the configuration

$$(G_1, \cdots, G_N, p_1, \cdots, p_{M_N}, d_1, \cdots, d_{M_N}) \notin \Delta(B_1, \cdots, B_S),$$

that

$$\prod_{i=1}^{N} \prod_{j=M_{i-1}+1}^{M_i} A_{ji}(p_j, G_i) \times H(p_j, D_j) \times \prod_{t=1}^{S} Q_t(D_{t_1}, \cdots, D_{t_{k(t)}})$$

$$> \prod_{i=1}^{N} \prod_{j=M_{i-1}+1}^{M_i} A_{ji}(p_j, G_i) \times H(p_j, d_j)$$

$$\times \prod_{t=1}^{S} Q_t(d_{t_1}, \cdots, d_{t_{k(t)}});$$

hence, the maximizing configuration lies in $\Delta(B_1, \cdots, B_S)$. *Q.E.D. of Lemma 1*

Our problem is therefore to maximize Equation (6) subject to $\mathbb{C} \in \Delta(B_1, \cdots, B_S)$. By the multiplication rule of conditional probability, Equation (6) can be written as

$$f[G_1, \cdots, G_N, p_1, \cdots, p_{M_N}, D_1, \cdots, D_{M_N} | B_1, \cdots, B_S]$$

$$= f[G_1, \cdots, G_N, p_1, \cdots, p_{M_N} | D_1, \cdots, D_{M_N}, B_1, \cdots, B_S]$$

$$\times f[D_1, \cdots, D_{M_N} | B_1, \cdots, B_S], \tag{A1}$$

which by the Markov property (4) is equal to

$$f[G_1, \cdots, G_N, p_1, \cdots, p_{M_N} | D_1, \cdots, D_{M_N}]$$

$$\times f[D_1, \cdots, D_{M_N} | B_1, \cdots, B_S]. \tag{A2}$$

The first term in (A2) is equivalent to

$$\tilde{f}[G_1, \cdots, G_N, p_1, \cdots, p_{M_N} | D_1, \cdots, D_{M_N}],$$

and if we assume $\mathbb{C} \in \Delta(B_1, \cdots, B_S)$, the second term can be written as

$$\frac{\tilde{f}[D_1, \cdots, D_{M_N}] \times \tilde{Z} \times \prod_{t=1}^{S} q_t(B_t)}{W},$$

where

$$W = \int_{D_1} \cdots \int_{D_{M_N}} \sum_{G_1, \cdots, G_N} \sum_{p_1, \cdots, p_{M_N}} \left\{ \prod_{i=1}^{N} \prod_{j=M_{i-1}+1}^{M_i} A_{ji}(p_j, G_i) \right.$$

$$\left. \times H(p_j, D_j) \times \prod_{t=1}^{S} Q_t(D_{t_1}, \cdots, D_{t_{k(t)}}, B_t) \right\} dD_1, \cdots, dD_{M_N}.$$

Therefore, for $\mathbb{C} \in \Delta(B_1, \cdots, B_S)$, Equation (6) can be written as

$$f[G_1, \cdots, G_N, p_1, \cdots, p_{M_N}, D_1, \cdots, D_{M_N} | B_1, \cdots, B_S]$$

$$= \tilde{f}[G_1, \cdots, G_N, p_1, \cdots, p_{M_N}, D_1, \cdots, D_{M_N}]$$

$$\times K(B_1, \cdots, B_S). \tag{A3}$$

Equation (A3) implies that $\mathbb{C}_{\max}(\mathbb{P}) = \tilde{\mathbb{C}}_{\max}(\Delta)$. *Q.E.D. of Theorem 1*

Proof of Theorem 2 With the structure of $f[D_j | p_j]$ as defined in (10), the optimal value of D_j minimizes

$$\sum_{j=t_1}^{t_{k(t)}} (D_j - \omega_j)^2 \{D_j \neq 0\}$$

subject to Equation (9). The result is found by using Lagrange multipliers. *Q.E.D. of Theorem 2*

Proof of Theorem 3 Define the following probability measure:

$$\Phi(\mathbf{G}) = \Phi(G_1, \cdots, G_N)$$

$$= \frac{1}{Z_\Phi} \prod_{i=1}^{N} \prod_{j=M_{i-1}+1}^{M_i} f[D_j(\mathbf{G}) | p_j(\mathbf{G})]$$

$$\times P[p_j(\mathbf{G}) | G_{(j)}] \times P[G_i]$$

$$= \frac{1}{Z_\Phi} \prod_{j=1}^{M_N} f[D_j(\mathbf{G}) | p_j(\mathbf{G})]$$

$$\times P[p_j(\mathbf{G}) | G_{(j)}] \times \prod_{i=1}^{N} P[G_i], \tag{A4}$$

where

$$Z_\Phi = \sum_{G_1} \cdots \sum_{G_N} \prod_{j=1}^{M_N} f[D_j(\mathbf{G})|p_j(\mathbf{G})] \times P[p_j(\mathbf{G})|G_{(j)}] \times \prod_{i=1}^{N} P[G_i].$$

Since $\Phi(\mathbf{G}) > 0 \; \forall \; \mathbf{G}$, Equation (A4) defines a new Markov random field on the G sites. It follows from Equation (17) that the most probable configuration of the G sites under $\Phi(\mathbf{G})$ (with corresponding p and D sites) is the most probable configuration of the graph whose probability density function is given by (5b).

The neighborhood structure of this new Markov random field is given by the set $N(i)$. This becomes apparent if we rewrite Equation (A4) as follows:

$$\begin{aligned}\Phi(G_1, \cdots, G_N) = \frac{1}{Z_\Phi} & \prod_{j \in \Pi_i \cup \Omega_i} f[D_j(\mathbf{G})|p_j(\mathbf{G})] \\ & \times P[p_j(\mathbf{G})|G_{(j)}] \times \prod_{G \in N(i)} P[G] \\ & \times \prod_{j \notin \Pi_i \cup \Omega_i} f[D_j(\mathbf{G})|p_j(\mathbf{G})] \\ & \times P[p_j(\mathbf{G})|G_{(j)}] \times \prod_{G \notin N(i)} P[G], \qquad \text{(A5)}\end{aligned}$$

and note that the conditional probability distribution is given by Equation (18).

We note that the updating algorithm described above is nothing more than the stochastic relaxation algorithm on this new Markov random field. When this algorithm is applied, the configuration which occurs most frequently coincides with the most probable configuration as given by (17). Since this value for $\mathbf{G}$ with the associated p and D sites is $\bar{\mathbb{C}}_{\max}(\Delta)$, which by Theorem 1 equals $\mathbb{C}_{\max}(\mathbb{P})$, Theorem 3 is proved. *Q.E.D. of Theorem 3*

Acknowledgments

The author thanks George Gifford from IBM East Fishkill for suggesting the problem of interpreting OES data and for supplying the table values and data used for model construction and analysis. This work would not have been possible without his guidance and valuable insights. Our stimulating conversations and his many valuable suggestions continue to be an important resource for this research. The author is also grateful for the support provided for this research by Ben Hsiao, Bob Doxtator, Doug Bossen, Gary Behm, Bill Heller, and Richard Talbot, and for their helpful criticisms and suggestions. In addition, the author would like to express his appreciation to Ron Ericson, who, through his C programming expertise, was able to identify inefficiencies in the coding and thereby cut the processing time considerably. The author is indebted to the editor and the reviewers for their valuable comments and suggestions.

References

1. D. Montgomery, *Introduction to Statistical Quality Control*, John Wiley, New York, 1991.
2. G. Gifford, "Applications of Optical Emission Spectroscopy in Plasma Manufacturing Systems," *Proceedings of the SPIE Microelectronic Integrated Processing Symposium*, 1990, Vol. 1392–40, pp. 454–465.
3. J. Shabushnig, P. Demko, and R. Savage, "Applications of Optical Emission Spectroscopy to Semiconductor Processing," *Spectrosc.* **2**, 40–42 (1987).
4. M. Splichal and H. Anderson, "Application of Chemometrics to Optical Emission Spectroscopy for Plasma Monitoring," *Proc. SPIE* **1594**, 189–203 (1992).
5. J. Reader and C. Corliss, "Line Spectra of the Elements," *CRC Handbook of Chemistry and Physics*, CRC Press Inc., Boca Raton, 1989, p. E-249.
6. L. Powell, "Computer Search and Identification of Infrared Spectra by Correlation Techniques," Research Thesis, Indiana University, Bloomington, 1977.
7. P. Klahr and D. Waterman, *Expert Systems Techniques, Tools and Applications*, Addison Wesley Publishing Co., Reading, MA, 1986.
8. B. Soucek and M. Soucek, *Neural and Massively Parallel Computers: the Sixth Generation*, John Wiley, New York, 1988.
9. R. Duda and P. Hart, *Pattern Classification and Scene Analysis*, John Wiley, New York, 1973.
10. U. Grenander, *Lectures in Pattern Theory*, Vols. I, II, and III, Springer-Verlag, New York, 1976, 1978, 1981.
11. U. Grenander, "The Rietz Lecture 1985: Advances in Pattern Theory," *Ann. Statist.* **17**, 1–30 (1989).
12. R. Kindermann and J. Snell, *Markov Random Fields and their Applications*, American Mathematical Society, Providence, RI, 1980.
13. J. Besag, "Spatial Interaction and the Statistical Analysis of Lattice Systems (with discussion)," *J. Roy. Statist. Soc.* **36**, 192–236 (1974).
14. S. Geman and D. Geman, "Stochastic Relaxation, Gibbs Distributions and the Bayesian Restoration of Images," *IEEE Trans. Pattern Anal. Machine Intell.* **PAMI-6**, 721–741 (1984).
15. U. Grenander, *Abstract Inference*, John Wiley, New York, 1981.

Received April 16, 1992; accepted for publication April 8, 1993

Brock E. Osborn *IBM Enterprise Systems, 522 South Road, Poughkeepsie, New York 12601 (BROCK at TDCSYS3).* Dr. Osborn is an applied mathematician in the Fault Tolerance and Technology Laboratory. He joined IBM in 1977 as a systems programmer and received his Ph.D. degree in applied mathematics in 1986 from Brown University while on an educational leave of absence. Currently Dr. Osborn does

statistical consulting for four IBM business units: Enterprise Systems, Technology Products, Application Solutions, and Personal Systems. His research interests include stochastic modeling, reliability and quality, pattern recognition, and mathematical modeling in industrial applications. Dr. Osborn was elected into the New York Academy of Sciences in 1989; he is the vice president of the Mid-Hudson chapter of the American Statistical Association. He is a member of Sigma Xi, the Mathematical Association of America, and the Institute of Mathematical Statistics.