

Optimal reduction of a monitoring grid for SiO₂ deposition surface over a wafer for semiconductor devices

Riccardo Borgoni, Luigi Radaelli, Valeria Tritto, Diego Zappa

Abstract Semiconductor devices are used in all electronic equipment. These devices, commonly known as integrated circuit (IC) chips, are developed over the surface of a disk, called wafer, after deposition of a thin silicon substrate. The deposition of a SiO₂ film on the top of the wafer is used as interconnection and isolation of different layers. Film deposition is critical to further steps of IC manufacturing. For this reason thickness is monitored during the manufacturing process in correspondence of a grid of points defined on the wafer surface. Considering that collecting thickness measures is quite expensive, this paper addresses the problem of determining whether, given previous measures of the SiO₂ film on the in-use map, a new sub map of the initial one may be adopted minimizing a suitable loss functions of prediction. This paper presents a strategy to reduce the monitoring grid based on a spatial prediction algorithm.

Keywords: Statistics for business and industry, spatial sampling, simulated annealing, statistics for microelectronics

¹

Riccardo Borgoni, Department of Statistics, University of Milano-Bicocca;
riccardo.borgoni@unimib.it

Luigi Radaelli, SPC & Robustness Group, Numonyx Italy;
Luigi.RADAELLI@numonyx.com

Valeria Tritto, IRER Istituto Regionale di Ricerca della Lombardia, Milano;
valesky85@msn.com

Diego Zappa, Department of Statistical Sciences, Catholic University of Milan;
diego.zappa@unicatt.it

1 Introduction

Integrated circuits (IC) are an essential part of our day-by-day life. Their fabrication process consists of a sequence of several different physical and chemical steps performed on a thin silicon slice, called wafer. Among the different process steps, the silicon dioxide (SiO₂) deposition covers a crucial role, being the silicon dioxide used both as a passive element (i.e.: electrical insulation between two electrically active elemental components of the IC) and as an active element. For the production process considered in this paper the SiO₂ thickness has a target value of 1800 Angstrom over all the wafer surface. However, this value is not constant over the wafer surface but it can randomly deviate from the target due to structural causes. In order to control the deposition process and ensure optimal IC performances, the SiO₂ layer thickness across the wafer and across the boat must be carefully evaluated. For this reason, maps containing wafer coordinates where thickness must be measured are available. Assessing whether the deposited film thickness is uniform (or almost uniform) over the wafer surface is essential for further steps in chip manufacturing. Data collection procedures are time consuming and expensive: for this reason it is often worth trying to reduce the number of points that are necessary to accurately reproduce the film thickness over the wafer surface.

Careful design of the sampling scheme can save time and money, and spatial statistics may be used to optimize such a scheme. Optimizing spatial sampling designs is important, particularly when data collection must be repeated many times over extended periods. Defining an optimal spatial design usually involves the minimization of a criterion based on a suitable function often related to stochastic algorithms used for prediction. An example is the kriging (Spruill and Candela, 1990).

Reducing a spatial monitoring grid requires to select a subset of the original measurement points in such a way that the result is the best possible estimate of the variable of interest. The minimization problem of the prediction error has not a unique solution and mostly depends on what the experimenter wants to keep under control. The matter is that whatever the approach we use, when the number of measurement locations in the starting map is high, a formidable combinatorial problem arises. The solution must be obtained by an adequate optimization algorithm, which converges as fast as possible to a global optimum. In collaboration with Numonyx, a semiconductor manufacturing company, we have successfully applied the simulated annealing (SA) method (Aarts and Korst, 1990). SA was firstly employed in spatial sampling by Sacks and Schiller (1988) and extended by Van Groenigen and Stein (1998). It was recently employed by Di Zio et al (2004) in conjunction with a fractal-dimension-based criterion and by Fuentes et al. (2007) in an entropy-based Bayesian paradigm. Diggle and Lophaven (2006) also discussed the problem of reducing a monitoring grid from a Bayesian perspective.

This paper addresses the problem of determining the sample design of a map to evaluate the thickness of the SiO₂ deposited on a silicon wafer when previous measures of the silicon oxide film are available. More specifically, starting from an existing design to monitor thickness, a methodology to reduce the monitoring map according to an optimal criterion is discussed. The following section describes a design that should respond to the needs described earlier whereas section 3 shows an application to a case study. Conclusions are drawn in section 4.

2 Optimal Spatial Sampling Design

In this section, a method based on a geostatistical model is suggested to choose the set of sampling locations, called the *design*. This methodology has been applied in the next section for defining a sampling map for predicting SiO₂ surface.

2.1 The Kriging Paradigm

A key issue in spatial analysis is the prediction of a spatial surface $Y(x)$ over a region W , using measurements, Y_i : $i=1, \dots, n$, taken at a finite collection of sampling locations $x_1, \dots, x_n$, $x_i \in W$, $i=1, \dots, n$, called the design.

A widely used stochastic model for such data assumes that the spatial surface of interest is a realization of a spatial process $\{Y(x): x \in W \subset \mathbb{R}^2\}$. It is largely accepted that the structured spatial variability of $Y(x)$ can be decomposed in at least two components: a large and a small scale variation. This is realized by assuming an additive structure $Y(x) = \mu(x) + Z(x)$ where $\mu(x)$ is a deterministic drift function and $Z(x)$ is a zero mean stationary intrinsic random process with variogram $2\gamma(h) = \text{Var}(Z(x+h) - Z(x))$.

The kriging predictor of Y in x is the minimum mean square error linear predictor, $\hat{Y}(x) = E[Y(x) | y_1, \dots, y_n]$. This predictor is unbiased with mean square error $\sigma^2(x) = E[(\hat{Y}(x) - Y(x))^2]$, which is usually termed the kriging variance. The

kriging is called universal if the drift function is not constant and ordinary in the particular case of a constant mean function (Chilès and Delfiner 1999 pg 164-175).

The optimization of a sampling scheme for the prediction of $Y(x)$ can be based on various criteria. Since the kriging variance is independent from the data, it may be considered alike an unconditional variance. It can be computed for every desired spatial location once the variogram is known and hence it can be used to define the spatial design. In particular it provides a criterion either to update retrospectively a monitoring network i.e. for adding or removing sampling points from an in-use design, or in a perspective manner when one needs to design in advance the data collection. It is known that designs, built to be efficient for parameter estimation, are not necessarily so for spatial prediction given the model parameters (Müller, 2001). Because the motivating substantive problem sketched in the introduction, the aim of this paper is to define an optimal sampling design which is efficient for the spatial prediction of the surface of interest given the correlation structure. However, had in a particular application something other than spatial prediction been the scientific concern, the approach proposed here would be inappropriate.

2.2 Simulated Annealing For Spatial Sampling

Let M be a finite set of points and $f: M^n \rightarrow \mathbb{R}_+$ be a positive function, called the fitness function, to be optimized over M^n . Simulating annealing (Aarts and Korst, 1990) is an

algorithm for optimizing such a function when the number of possible configurations of values of M is extremely high and cannot be evaluated by enumeration.

The algorithm starts from a configuration of points, say $S_0 \subseteq M^n$ and sequentially updates it. At each step i , the current configuration S_i is modified by replacing one point by one point included in its complement $\bar{S}_i$. The candidate point for replacement is selected randomly in a neighbourhood of the current sample and it is accepted if this determines an improvement in the value of the fitness. Otherwise the decision whether the candidate should be included is randomized. This implies that the algorithm allows for configurations less favourable than the current one. The probability of taking such less favourable subset of points, decreases as the algorithm goes on. The procedure is iterated until the value of the fitness gets stable and cannot be further reduced. The criterion used for updating the current sample, known as the Metropolis criterion, is

$$P(S_i \rightarrow S_{i+1}) = \begin{cases} 1 & \text{if } f(S_{i+1}) \leq f(S_i) \\ \exp\left(\frac{f(S_{i+1}) - f(S_i)}{c}\right) & \text{if } f(S_{i+1}) > f(S_i) \end{cases}$$

The arrow means replacement and c is a constant which decreases in value as the algorithm keeps going, lowering the probability of accepting less favourable configurations.

In the context of spatial sampling M is a lattice of possible measurement points in W and S_i is a subset of the lattice. The algorithm in the form described above is particularly suitable for retrospective spatial design where one has to decide which points must be removed from a monitoring network defined by a pre-existing lattice.

Aiming at optimizing the design for surface prediction the fitness function should be related to the kriging variance, evaluated at the current sample configuration. In this paper we have considered the fitness function proposed by van Groenigen et alii (1999) for ordinary kriging, namely $\sigma = \int_W \sigma^2(x|S) dx$. This is an integrated mean

square prediction error obtained by integrating the kriging variance over the region W of interest. The actual computation of the integral of interest has been worked out by superimposing a fine grid, G_W , of size n_G over the region W and approximating the integral by $(n_G)^{-1} \sum_{u \in G_W} \sigma^2(u|S)$.

3 A Case Study: Reducing the Monitoring Grid of SiO₂ Deposition on a Wafer

In this section the strategy described above has been applied to the problem of reducing the size of a grid used for monitoring the thickness of SiO₂ deposited on a wafer.

The dataset used refers to measurements taken on ten wafers using a map of 49 locations. One point of the grid is located at the centre. The others lie on three concentric circumferences of different radius as shown by figure 1-(a).

The aim is to narrow the size of the grid down to 20 points, a number believed adequate by the engineers. The sampling scheme used is stratified supposing that the

same proportion of points in each circumference inside the wafer in the reduced design must be equal to the proportion in the original grid. It was also decided to constraint the algorithm to retain the central point in the final grid to estimate as precisely as possible the valley of the deposition surface observed at the centre of the wafer, see figure 1-(a). The reason is that this is a peculiar characteristic of the deposition process.

By an explorative analysis two wafers considered anomalous both for the trend and the correlation structure were identified and initially discarded from the analysis. In fact it turned out that these wafers were processed by equipment different from the others. Figure 1-(b) and 1-(c) show respectively the surface trend and the empirical variogram of the eight wafers used in the analysis. The other wafers present similar shapes to the drift and variogram. From the variogram inspection it appears a clear effect of the trend on the correlation structure which determines a striking increase in the values of the variogram bins followed by a clear drop at larger distances.

Three randomly selected wafers were used as a training set to estimate the parameters of the variogram and trend surface; the remaining 5 were used for validation purposes. Figure 2-(a) shows the maximum likelihood estimate of the exponential variogram for the analyzed data. Because of the trend structure, universal kriging was used to predict the deposition surface assuming a complete polynomial of second order for the mean function of the process.

SA algorithm was applied to one wafer randomly selected amongst the three used to train the procedure. The algorithm started from a random sample of 20 locations taken without replacement from the initial 49 points. Convergence was reached after 10041 iterations, the acceptance rate being 19%. Figure 2-(b) shows the shape of the fitness for the retained configurations throughout the iterations performed. The optimal map obtained at the end of the procedure is reported in Figure 2-(c), along with the points removed from the design. This map reduces the fitness value by about 17% with respect to the starting configuration.

Figure 1: (a) Monitoring grid for thickness deposition; (b) local polynomial trend of deposition surface; (c) the empirical variogram estimated

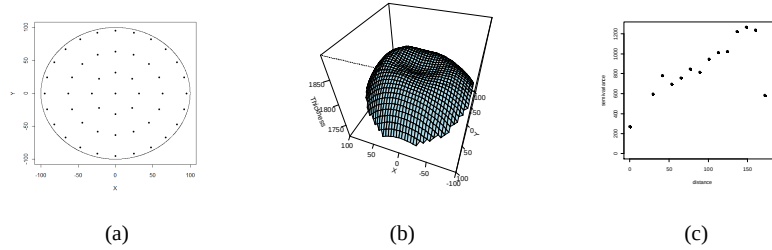


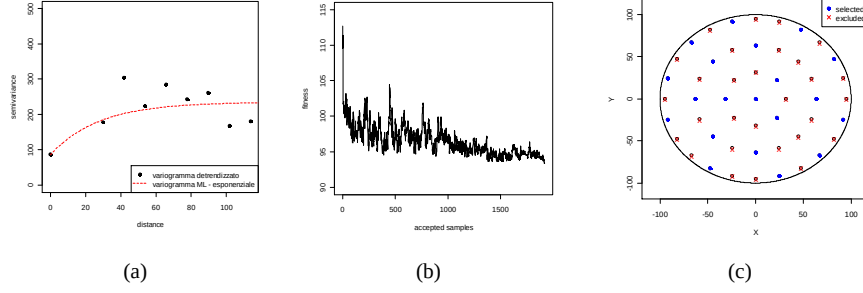
Figure 3 shows the estimated deposition surface by using the complete 49-points grid (Figure 3-a) and the reduced 20-points grid S (Figure 3-b). It appears that the reduced sample manages to grasp salient characteristics of the surface in a manner which is comparable to the complete one. Figure 3-(c) shows that a reasonable good fit was obtained when predicting the 29 points left outside the design and also it shows that larger residuals tend to cluster on the external circumference of S , see figure 3-(d). In order to further evaluate the final design, the grid was crossvalidated by the seven wafers not used for setting it. The two anomalous wafers were also considered for comparative purposes. This was done by computing the average prediction error over the 29 location in $\bar{S}$

Optimal reduction of a monitoring grid for SiO₂ deposition surface over a semiconductor wafer 6

$$\frac{1}{\# \bar{S}} \sum_{i \in \bar{S}} E \left[\hat{Y}_i - Y_i \right]^2 \quad (1)$$

where $\hat{Y}_i$ and Y_i , are the predicted and the actual measures of deposition at site i .

Figure 2: (a) Estimated exponential variogram: points represent the empirical variogram calculated by residuals; (b) SA fitness function values for the accepted sampling configuration; (c) final monitoring map represented by dots superimposed to the monitoring grid in use



The value of (1) ranged from 112 to 174 in the seven wafers not used for kriging the deposition surface, whereas it equals 125 for the wafer used for setting the grid. Similar values of (1) were also found for the two anomalous wafers (127 and 117).

To assess the hypothesis on whether the deposition model estimated on the selected wafer is able to reproduce what was observed in other wafers, a multigaussian model was used. More specifically a Gaussian process was simulated for each point excluded from the design via sequential conditional simulations, using the model estimated above. The process was simulated one thousand times on each of the 29 points in $\bar{S}$. The maximum and the minimum value generated for each grid point were considered for building an envelope which can be considered consistent with the sampling variability of the hypothesized deposition model. The values actually observed for each points in $\bar{S}$ were compared with the generated envelope.

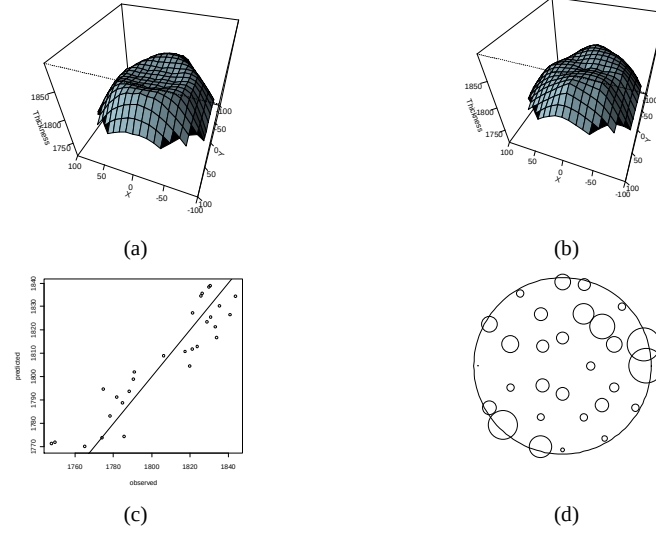
The multigaussian model seems to be adequate for the wafer used for reducing the design as only one measure originally observed on $\bar{S}$ was outside the envelope. The result for all the remaining 9 wafers is reported in figure 4.

The percentage of points laying outside the envelope was about 18%. This percentage is much lower for the 7 wafers not identified as anomalous in the explorative analysis. In this case the number of points where the prediction envelope did not contain the actual values ranged from 0 to 8. This number was much higher for the two anomalous wafers where 14 and 16 points were observed outside the envelope. This was clearly an expected result if one considers what was said above. However this also suggests an additional use of the procedure outlined in this paper as an instrument which can be employed for assessing the quality of the deposition process.

Once a model which is believed adequate for deposition, for instance on an experience basis, has been estimated, this model can be used to generate an envelope as explained above. A high proportion of measures taken on a new wafer not included in the envelope can suggest potential problems in the manufacturing process.

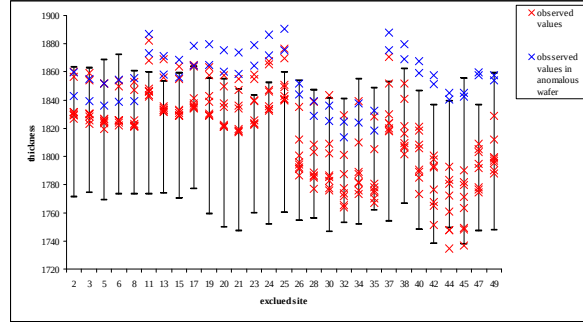
Optimal reduction of a monitoring grid for SiO₂ deposition surface over a semiconductor wafer 7

Figure 3: (a) Universal kriging prediction obtained by the complete grid; (b) the reduced grid; (c) predicted versus observed values; (d) and residuals $|Y_i - \hat{Y}_i|$ on points removed from the design



A map, similar the one showed in figure 4-(d), can identify the regions of the wafer where the problem tends to occur.

Figure 4: Simulated deposition envelope for the points removed from the design and actual values observed in each wafer. The id number of each point is reported in the abscissa axis.



4 Conclusions

In this paper, we considered an approach based on spatial statistics to optimal reduction of a spatial monitoring grid for SiO₂ deposition film on a wafer surface for semiconductor devices. The optimality was established in terms of the mean squared

error of the spatial prediction. The combinatory nature of the problem was tackled using the simulated annealing algorithm. Such an algorithm is flexible enough to handle the constraints requested by the sample scheme, once these constraints can be translated into well defined geometric structures. In particular this makes possible to adopt a stratified design and to force some preselected points of the original map in the reduced sample. Furthermore the proposed methodology can be combined with other sampling schemes, fitness function and spatial prediction methods. Although we cannot report the results here, a simple random sample was compared to the stratified one obtaining similar maps with an overall worse performance in terms of the fitness. It can be noted that if one plans to set a regular monitoring grid, optimal designs are known for the kriging prediction of a stationary random field (Yfantis et al. 1987). However, this strategy cannot be always employed, as it is the case of the case study presented in this paper where the measurement locations are fixed by technical constraints.

A more conservative fitness function was also considered e.g. $\max_{u \in G_w} \sigma^2(u | S)$. It was observed that this fitness function sampled much more the external circumference if combined with a simple random sampling scheme. In other words it tended to protect towards the occurrence of large prediction errors which cluster on such circumference, whilst it showed a worse average performance than the fitness function in (1). Finally we'd like to just mention that a more robust approach to spatial prediction can be used within the proposed methodology to account for large scale variability such as the local kriging which, in some situations, can also have computational advantages. However this approach was not worth in the present context because the relative small size of the dataset involved and the clear structure of the process drift can be easily and more powerfully grasped parametrically.

References

1. Aarts, E., Korst, J.: Simulated Annealing and Boltzmann Machines - A Stochastic Approach to Combinatorial Optimization and Neural Computing,. Wiley, New York (1990)
2. Chiles, J.P., Delfiner, P.: Geostatistics: Modeling Spatial Uncertainty, John Wiley & Sons, New York (1999)
3. Di Zio S., Fontanella L., Ippoliti L.: Optimal Spatial Sampling Schemes for Environmental Surveys. Environmental and Ecological Statistics, 11, 4, 397-414, (2004).
4. Diggle, P., Lophaven, S.: Bayesian Geostatistical Design, Scandinavian Journal of Statistics, 33: 53-64 (2006)
5. Fuentes M., Chaudhuri A., Holland D.M.: Bayesian entropy for spatial sampling design of environmental data, Environ Ecol Stat 14:323-340 (2007)
6. Müller, W. G.: Collecting spatial data, Springer-Verlag, Heidelberg (2001)
7. Sacks, J., Schiller, S.: Spatial designs. In: Gupta, S.S., Berger, J.O. (eds.), Statistical Decision Theory and Related Topics IV, 2, pp. 385-399. Springer-Verlag, New York (1988)
8. Spruill, T., Candela L.: Two approaches to design of monitoring networks, Ground Water, 28, 430-442 (1990)
9. van Groenigen, J.W., Siderius, W., Stein, A.: Constrained optimisation of soil sampling for minimisation of the kriging variance Geoderma, 87, 239-259 (1999)
10. van Groenigen, J.W., Stein A.: Constrained optimization using continuous simulated annealing, Journal of Environmental Quality, 27, 1078-1086 (1998)
11. Yfantis E.A., Flatman G.T., Behar, J.V.: Efficiency of kriging estimation for square, triangular, and hexagonal grids. Mathematical Geology, 19:183-205 (1987)