Non-parametric local capability indices for industrial planar artefacts

Riccardo Borgoni
Università di Milano-Bicocca
Milan Italy

May 18, 2021

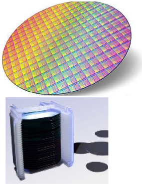
ENBIS 2021 Spring Meeting

Joint work with Diego Zappa and Vincenzo Emanuele Farace

Outline

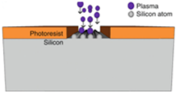
- motivating problem
- local process capability indices
- review of quantile and additive quantile regression
- numerical evidences
- case study

Etching in the Integrated Circuit (IC) fabrication process



The IC fabrication process consists of a sequence of physical and chemical steps performed on a circular thin silicon slice, called a wafer (W).
Wafers are typically processed in lots.

During the process, silicon dioxide thin films are deposited on the wafer surface mostly as an electrical insulation between two electrically active components of the chip.



To this end, a pattern of trenches is etched on the silicon substratum.

Trench depth is expected to conform to specification limits and depth variability with respect to these limits is typically evaluated by appropriate indices

Process Capability Indeces (PCI)

Y: a random variable representing the trench depth

PCI are statistics that measure the ability of an in-control process to give values of Y within specification limits, hence representing the *natural variation* of the production process

C_{pk} is a capability index routinely adopted in practice.
It can be calculated (Clements, 1989) by

$$C_{pk} = min(C_{pu}, C_{pl}) \quad C_{pu} = \frac{USL - \xi_{0.5}}{\xi_{0.9986} - \xi_{0.5}} \text{ and } C_{pl} = \frac{\xi_{0.5} - LSL}{\xi_{0.5} - \xi_{0.00135}}$$

- LSL: lower specification limit
- USL: upper specification limit
- ξ_τ: τ-quantile of Y, $\tau \in (0,1)$

Process Capability Indeces contd...)

- C_{pk} calculation requires many observations to estimate the required quantiles with appropriate precision
- C_{pk} are often computed at the lot level pulling together measures taken at different locations of the wafer and in different wafers of a given lot
- In planar manufacts, it can be relevant to assess process capability locally. In particular when the overall product surface is parcelled into separate pieces to realise different items. This is, in fact, the case of microchip production since hundreds or thousands of dice are obtained from each single wafer

Local (spatial) Process Capability Indeces

$\{Y(\boldsymbol{s}),\, \boldsymbol{s} \in W\}$: random field representing the output measured at different spatial points of the wafer W

A local version of C_{pk} can be defined by

$$C_{pk}(\boldsymbol{s}) = min\{C_{pu}(\boldsymbol{s}), C_{pl}(\boldsymbol{s})\}$$

$$C_{pu}(\boldsymbol{s}) = \frac{USL - \xi_{0.5}(\boldsymbol{s})}{\xi_{0.9986}(\boldsymbol{s}) - \xi_{0.5}(\boldsymbol{s})} \quad \text{and} \quad C_{pl}(\boldsymbol{s}) = \frac{\xi_{0.5}(\boldsymbol{s}) - LSL}{\xi_{0.5}(\boldsymbol{s}) - \xi_{0.00135}(\boldsymbol{s})}$$

$\xi_{\tau}(\boldsymbol{s})$: τ-quantile of Y at location $\boldsymbol{s} \in W$, $\tau \in (0, 1)$

In practice it is impossible to collect enough data to calculate $\xi_{\tau}(\boldsymbol{s})$ for each $\boldsymbol{s} \in W$ and some smoothing is required to provide a C_{pk} spatial surface

Mean, Median, Quantiles and Optimization

Y: (absolutely continuous) random variable with c.d.f. $F(y)$

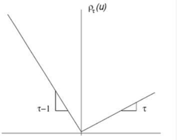
m: scalar

- Expectation: $\mathrm{E}(Y) = \mathrm{argmin}_m \, \mathrm{E}[(Y-m)^2]$

  loss function ℓ_2: $\rho(Y-m) = (Y-m)^2$
- τ-quantile: $\xi_\tau : F^{-1}(\xi_\tau) = \tau$

  $\xi_\tau = \mathrm{argmin}_m \, \mathrm{E}(\rho_\tau(Y-m)),\ \tau \in (0,1)$

  loss function $\rho_\tau(u) = \begin{cases} \tau u & \text{if } u > 0 \\ (\tau-1)u & \text{if } u \leq 0 \end{cases}$



$\tau = 0.5$ gives the median (ℓ_1 loss function)

Conditional Mean, Quantiles and Quantile Regression (QR)

x: p-vector of covariates.

By replacing m with $x'\boldsymbol{\beta}$

- multiple regression $\rho_\tau(u) = \rho(u) = ||u||^2$

$$\mathrm{E}(Y|\mathrm{x};\beta) = \mu(\mathrm{x}) = \mathrm{x}'\boldsymbol{\beta}$$

- quantile regression $\rho_\tau(u) = (\tau - I(u \leq 0))u$ (Koenker and Bassett, 1979)

$$Q_\tau(Y|\mathrm{x};\beta_\tau) = \xi_\tau(\mathrm{x}) = \mathrm{x}'\boldsymbol{\beta}_\tau$$

- $\tau = 0.5$: median regression
- $\boldsymbol{\beta}_\tau$ depends on the quantile order, hence a separate quantile model is obtained for each τ
  pause
- x may incorporate binary variables as well as polynomial

Parameter Estimation

$((y_i, \mathrm{x}_i),\ i = 1, \ldots, n)$: sample

- loss function ℓ_2: $\rho_\tau(u) = \rho(u) = ||u||^2 \rightarrow$ OLS

$$\hat{\beta} = \operatorname*{argmin}_{\beta} \sum_{i=1}^{n} (Y_i - \mathrm{x}_i' \beta)^2$$

- loss function $\rho_\tau(u) = (\tau - I(u \leq 0))u$

$$\hat{\beta}_\tau = \operatorname*{argmin}_{\beta} \sum_{i=1}^{n} \rho_\tau(Y_i - \mathrm{x}_i' \beta_\tau)$$

Linear programming algorithms are used to fit the model

General remarks

- Non-parametric
- Robustness
- Additive extensions are simply encompassed by this framework

Quantile Generalised Additive Model (qgam)

Quantile spatial field $\xi_\tau(\boldsymbol{s})$ is approximated by a bivariate spline

$$S_\tau(\boldsymbol{s}) = \sum_{j=1}^{k} B_j(\boldsymbol{s})\beta_{j,\tau}$$

- $B_j(\boldsymbol{s})\, j = 1, \ldots, k$ is a (known) bivariate thin plate basis function
- $\beta_{j,\tau}\ j = 1, \ldots, k$ are unknown coefficients

The τ-quantile spatial surface can be estimated at location $\boldsymbol{s}$ by

$$\hat{\xi}_\tau(\boldsymbol{s}) = \hat{S}_\tau(\boldsymbol{s}) = \sum_{j=1}^{k} B_j(\boldsymbol{s})\hat{\beta}_{j,\tau}$$

Bayesian qgam

Loss function: extended log-f (ELF)

$$\tilde{\rho}_\tau(u) = (\tau - 1)\frac{u}{\sigma} + \lambda \log\left(1 + e^{\frac{u}{\lambda\sigma}}\right)$$

- $\sigma > 0$: scale parameter
  $\sigma_0\sigma(x)$ in case of relevant covariates x for the scale
- $\lambda > 0$: tuning parameter

An improper $N(0, \Omega)$ is used for $\boldsymbol{\beta}$. Ω is p.s.d. precision matrix depending on smoothing parameters $\boldsymbol{\gamma}$.

Bayesian qgam - estimation

The approach is based on the loss function (no probabilistic model for the response)

Algorithms are used to performing the Bayesian update under the ELF loss using a loss-based pseudolikelihood (Bissiri et al. 2019, Fasiolo et al. 2020).

The estimation procedure is composed by several routines

- σ_0 is selected by minimizing a calibration loss function numerically
- λ and $\sigma(x)$ are estimated using close-form expressions
- smoothing parameters γ are selected by numerically optimizing an intermediate criterion
- $\boldsymbol{\beta}$ are obtained using maximum a posteriori (MAP) method keeping fixed the other parameters

Spatial C_{pk} estimation via qgam

The C_{pk} surface can be by computing

$$\hat{C}_{pk}(\mathrm{u}) = min\{\hat{C}_{pu}(\mathrm{u}), \hat{C}_{pl}(\mathrm{u})\}$$

$$\hat{C}_{pu}(\mathrm{u}) = \frac{USL - \hat{\xi}_{0.5}(\mathrm{u})}{\hat{\xi}_{0.9986}(\mathrm{u}) - \hat{\xi}_{0.5}(\mathrm{u})} \quad \text{and} \quad \hat{C}_{pl}(\mathrm{u}) = \frac{\hat{\xi}_{0.5}(\mathrm{u}) - LSL}{\hat{\xi}_{0.5}(\mathrm{u}) - \hat{\xi}_{0.00135}(\mathrm{u})}$$

over a fine grid of locations $G = \{\mathrm{u}_1, \dots \mathrm{u}_N : \mathrm{u}_j \in W\}$ prefixed on the wafer area

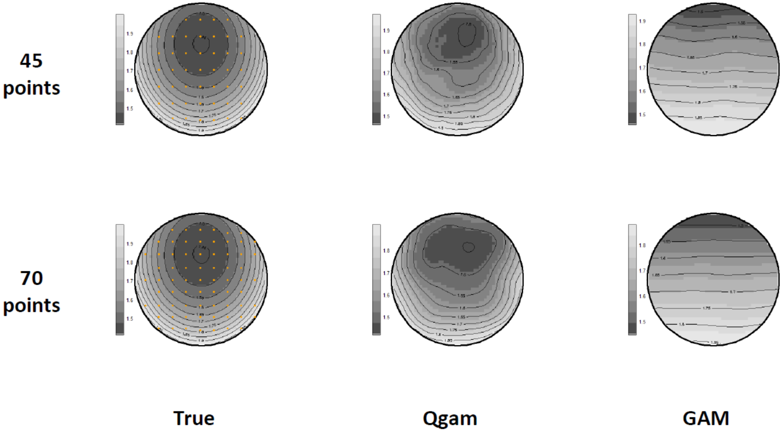
Numerical evidences

Scenario 1

- Simulation model: $Y(\boldsymbol{s}) \sim LN(\mu(\boldsymbol{s}), \sigma(\boldsymbol{s}))$
- $\mu(\boldsymbol{s})$ and $\sigma(\boldsymbol{s})$ are separately estimated from the lot data
- *True* $C_{pk}(\boldsymbol{s})$ surface is obtained using actual USL and LSL of the etching process
- Data are simulated at sampling monitoring grid of different size: 45 and 70 points
- Simulated data are used to estimate $\xi(\mathrm{u})$ via qgam over a prediction grid of 1876 points internal to the wafer area
- The exercise is repeated B=500 times and $\hat{C}^b_{pk}(\mathrm{u})$ is computed for each $\mathrm{u} \in G$ and $b = 1, \ldots, B$
- the average $\bar{C}_{pk}(\mathrm{u}) = B^{-1} \sum_{b=1}^{B} \hat{C}^b_{pk}(\mathrm{u})$ is computed for each grid point u

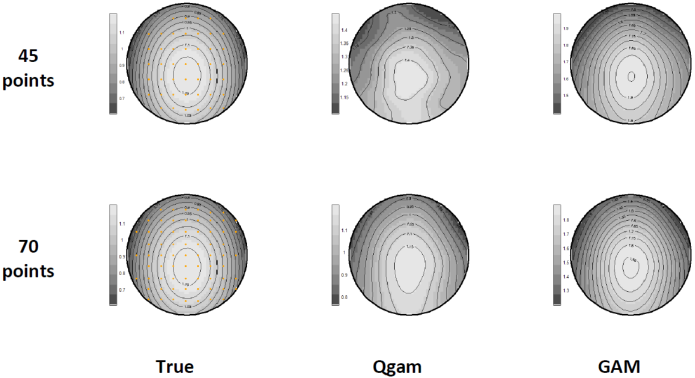
Results are benchmarked using the $C_{pk}(\boldsymbol{s})$ surface estimated via a semi-parametric additive model for log-normal data

Scenario 1: results



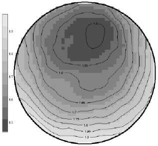
Scenario 2: results - robustness to model specification

Same as Scenario 1. Data simulated by a generalised T with df=9

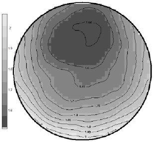


Scenario 3: results - robustness to outliers

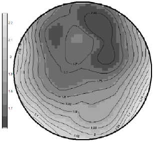
Same as Scenario 1. Data are perturbed forcing $\delta\%$ of outliers.
Outliers are generated by $Unif(a, b)$ where
a: minimum of the values generated by the LN scheme
b: 5th percentile of the values generated by the LN scheme



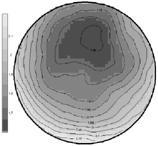
No outlier



Outliers: 5%



Outliers 10%

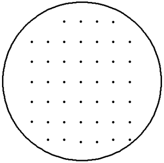


Outliers 15%

Assessing lot process capability: wafer data

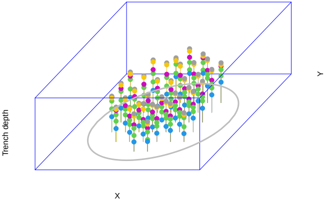
Notation

S: sample grid with $Card(S) = n = 38$



L: set of wafers in the lot

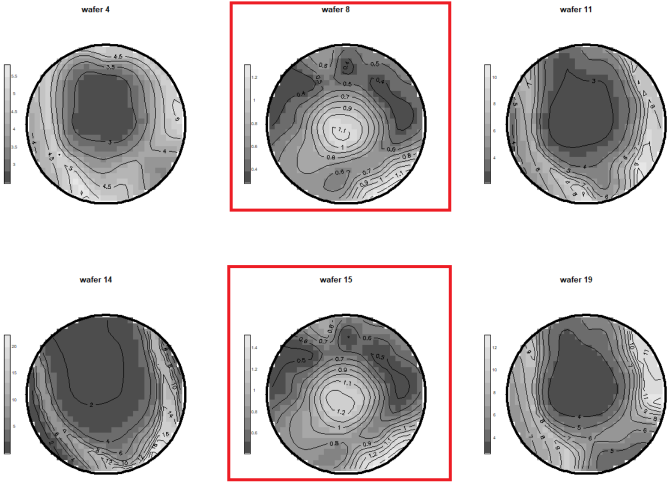
$M = 6$: number of wafers in the lot



G: prediction with $Card(G) = N = 648$

Assessing lot process capability: results - qgam C_{pk} surfaces

$\hat{C}_{pk}(\boldsymbol{s})$ surface of each wafer of the lot



Assessing lot process capability in wafer etching

- Process capability is typically assessed at the lot level.
- Wafers in the same lot may or may not be locally homogeneous in their capability
- Local homogeneity of the process capability has to be preliminary explored

Assessing lot process capability: procedure

❶ exclude wafer w from L and set $L' := L - w$

❷ estimate (qgam) $\hat{C}_{pk}^{w}(u)$ for each $\mathrm{u} \in G$ of wafer w

❸ for each $\boldsymbol{s} \in S$ extract randomly one value out of $M' = M - 1$ available and obtain a new sample $Y^* = \{Y^*(\boldsymbol{s}_1), \dots, Y^*(\boldsymbol{s}_n)\}$

❹ calculate $\hat{C}_{pk}^{*}(\mathrm{u})$ for each $\mathrm{u} \in G$ using Y^*

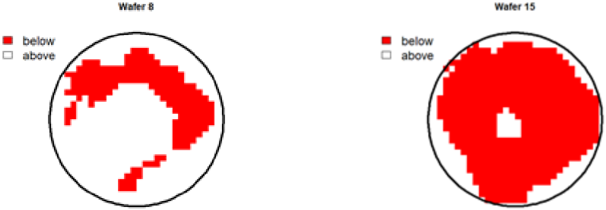
❺ repeat steps 3 and 4 B times, obtain $(\hat{C}_{pk}^{b*}(\mathrm{u}),\ b = 1, \dots, B)$ and calculate the 5%-percentile $C_{pk,0.05}^{*}(\mathrm{u})$ for each u

❻ Set $I(\mathrm{u}) = 1$ if $\hat{C}_{pk}^{w}(u) < C_{pk,0.05}^{*}(\mathrm{u})$ and 0 otherwise

❼ if $\sum_{\mathrm{u} \in G} I(\mathrm{u})/N > \delta = 0.20$ set $L := L'$, $M := M'$ and go to 1

Assessing lot process capability: results - capability map

The algorithm identifies two wafers with a "significantly lower" C_{pk} surface than "average"

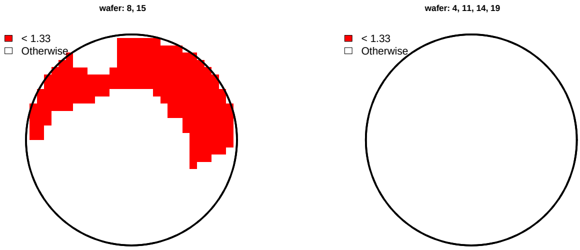


(a) Iteration 2: 28% of the wafer area below $C^*_{pk,0.05}$(u)

(b) Iteration 5: 72% of the wafer area below $C^*_{pk,0.05}$(u)

Assessing lot process capability: results - capability map

Process Capability per die in the two groups
Resampling is used to calculate a probability interval of local C_{pk} at each location $\boldsymbol{s} \in W$ (250 replicates)
Probability interval: 15th and 85th percentiles of $\hat{C}_{pk}(\boldsymbol{s})$



(c) Wafer 8,15: 33% dice with poor capability

(d) Wafer 4,11,14,19: 100% dice with excelent capability

Conclusions

Main points:

- a local estimate of the process capability based on additive quantile regression suitable for spatial (planar) measures
- a procedure based on this estimate to assess homogeneity of the process capability when items are produced in lots
- a procedure to assess local capability of the production process at the lot level

Bibliographic references

- Bissiri PG, Holmes CC, Walker SG (2016). A General Framework for Updating Belief Distributions. Journal of the Royal Statistical Society B, 78(5), 1103?1130
- Clements J, (1989) Process Capability Calculations for Non-Normal Distributions, Quality Progress, 22, 9, 95-100
- Fasiolo M, Wood SN, Zaffran M, Nedellec R, Goude Y (2020). Fast Calibrated Additive Quantile Regression. Journal of the Americal Statistical Association
- Koenker R, Bassett JG. (1978) Regression quantiles. Econometrica pp 33-50

Thank you
for your attention and comments