Received: 16 August 2021 | Revised: 3 January 2022 | Accepted: 25 January 2022

DOI: 10.1002/asmb.2673

RESEARCH ARTICLE

WILEY

Non-parametric local capability indices for industrial planar manufacts: An application to the etching phase in the microelectronic industry

Riccardo Borgoni[1] | **Vincenzo Emanuele Farace**[2] | **Diego Zappa**[3]

[1]Dipartimento di Economia, Metodi Quantitativi e Strategie d'Impresa, Università di Milano-Bicocca, Milan, Italy

[2]Università di Milano-Bicocca, Milan, Italy

[3]Dipartimento di Scienze Statistiche, Università Cattolica del Sacro Cuore, Milan, Italy

Correspondence
Riccardo Borgoni, Università di Milano-Bicocca, Milan, Italy.
Email: riccardo.borgoni@unimib.it

Funding information
Universita degli Studi di Milano-Bicocca

Abstract

Process capability indices are routinely used to estimate the mean-variability performance of industrial products with respect to both targets and specification limits. However, when the target variable is defined over a planar surface of a manufact, it is relevant to assess the capability of the production process locally, that is, at any spatial location of the surface, in particular if the manufact has to be split into pieces to obtain single production items. In this article, focusing on the C_{pk} specification introduced by Clements [Qual Prog., 22, 95–100], we suggest an approach based on additive quantile models to estimate, in a Bayesian paradigm, the index locally. We demonstrate its use in the context of the etching phase of the integrated circuit fabrication process. Since capability of etching processes is typically assessed for batches of wafers, we also propose two algorithms based on resampling to perform local capability analysis at the lot level.

KEYWORDS

Bayesian semiparametric quantile regression, dry etching, microelectronics, thin plate spline

1 | INTRODUCTION

Process capability indices (PCIs) are tools customarily adopted to estimate the mean-variability performance of industrial processes with respect to both targets and specification limits, and are nowadays a standard tool for commercial activity. PCIs used in industry are single numerical measures that summarise the process performance. De-Felipe and Benedito[1] provide a thorough review on this topic. Amongst different indicators of process capability, the C_{pk} index[2] has gained popularity. A production process is called capable if C_{pk} is above a certain threshold. A common reference value is 1.33.[1]

Hereinafter, we considered the C_{pk} specification introduced by Clements[3] that generalises the usual version of the index to account for potential asymmetry of the distribution of the target variable. Clements's quantile-based transformation approach has been fundamental for the development of process capability index estimators for potentially non-normal data following many different stochastic models (see Kotz and Lovelace[4] for a review).

In the case study discussed in this article, a fabrication process in semiconductor manufacturing, known as *dry etching*, is considered. The outcome of interest is the depth of trenches defined by photolithographic steps and etched into the wafer surface. During this phase of the integrated circuit fabrication process, the C_{pk} index is regularly computed to evaluate the

process capability of batches of wafers also termed *lots*. Given a lot of L wafers monitored using a grid of n points, the C_{pk} is often obtained in practice putting all the data together. The usual formula (see Equation (1)) of the index is used, ignoring any source of spatial variability and possibly only inspecting the data quality by adjusting for outliers presence. This approach would be correct in case data are from rational subgroups,[2] but barely this happens. In other words, a unique C_{pk} value is calculated for an entire lot to assess the process ability to produce the output within the specification limits.

Assumptions behind PCI calculation are that the outcome measurements are from an i.i.d. process and that the specification limits are known and fixed. In addition, if a target is known, it is supposed to be different from the specification limits. If data refer to batches of products, they are assumed to have the same quality amongst and within the lots.[5]

However, when considering planar manufacts, as in the context considered in this article, experience suggests that the process capability is rarely homogeneous over their entire surface and assessing it at a local level can be relevant. This is true, in particular, in those cases where the overall product surface has to be split into pieces to obtain the different items. This is, in fact, the case of microchip production since hundreds or even thousands of dice are obtained from each single wafer. In these circumstances, deriving a local version of the C_{pk} can provide relevant information about the production process at the chip level. Clearly the measurement effort that is necessary to locally monitor the production process even at a moderate spatial resolution is huge and some smoothing must be adopted to estimate the index surface. Despite planar manufacts being recurrent in many industrial production processes, this issue has not been thoroughly investigated.[6,7]

In this article, we take advantage of the quantile-based specification of Clements's index to propose a nonparametric approach grounded on additive quantile regression that permits us to estimate the process capability at any spatial location of interest in a fully nonparametric manner, that is, without assuming any specific stochastic model for the target variable.

Quantile regression was developed as an extension of the linear regression model in a seminal paper published by Koenker and Bassett in 1978. In this article, we adopt the Bayesian paradigm to additive quantile regression following the approach proposed by Fasiolo et al.[8] where bivariate splines are employed to grasp spatial regularities in the quantile surface.

As mentioned above, the capability of the etching process is typically assessed at the lot level. However, wafers in the same lot may or may not be homogeneous in terms of their local capability. To investigate wafer homogeneity, a procedure is proposed later on in the article that pairs quantile-regression-based C_{pk} estimates and resampling to identify subgroups of lot wafers that can be considered homogeneous. A second procedure is proposed to combine the information taken from a set of (capacity homogeneous) wafers to produce a local capability analysis at the lot level.

The article is organised as follows. In the next section, we introduce the motivating case study and describe the dataset at hand. Capability indices and local capability indices are considered in Section 3, whereas a brief review of quantile and additive quantile modelling, both in the classical and in the Bayesian context, is provided in Section 4. In Section 5, two algorithms are proposed to evaluate local process capacity for batches of wafers. Monte Carlo evidence of the performance of the proposed method is reported in Section 6, whereas we present an application of the proposed methodology in Section 7. Section 8 ends the article with some conclusions and final remarks.

2 | THE MOTIVATING CASE STUDY: ETCHING PROCESS AND TRENCH DATA

Semiconductor devices are developed by highly integrated sequences of several technological steps on circular-shaped working substrates called wafers. In the final production phase, wafers are cut into small items, called *dice*, and then packaged into chipsets (see Figure 1A). All the phases of the fabrication process are carefully monitored. Data are collected by a network of measurement points to assess process stability, and punctual measurements are used to make inferences on the response surface over the entire wafer area in order to check whether quality standards are met even where actual measures have not been collected.[6,7,9]

Hereinafter, we consider the dry-etching phase of the integrated circuit production process. One of the preliminary steps of this process consists in the deposition of a thin SiO_2 film over the wafers. Etching technologies aim to transfer the patterns defined by photolithographic steps to the surface of the wafers by selective material removal (Figure 1B). The current main option is plasma etching, also called dry etching, which allows precise control of the profile dimensions and structures.

Wafers are typically processed in batches, that is, the surfaces of a number of wafers are simultaneously exposed to the plasma radicals and ions that, reacting with the superficial material in a vacuum chamber, easily remove oxide producing the desired straight profiles.

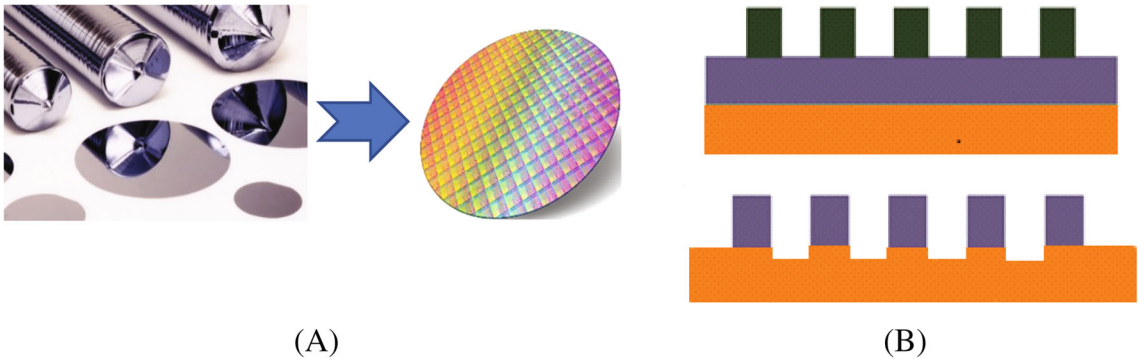

FIGURE 1 From virgin silicon wafer to chipset (A); Photoresist thickness before and after the etching process (B)

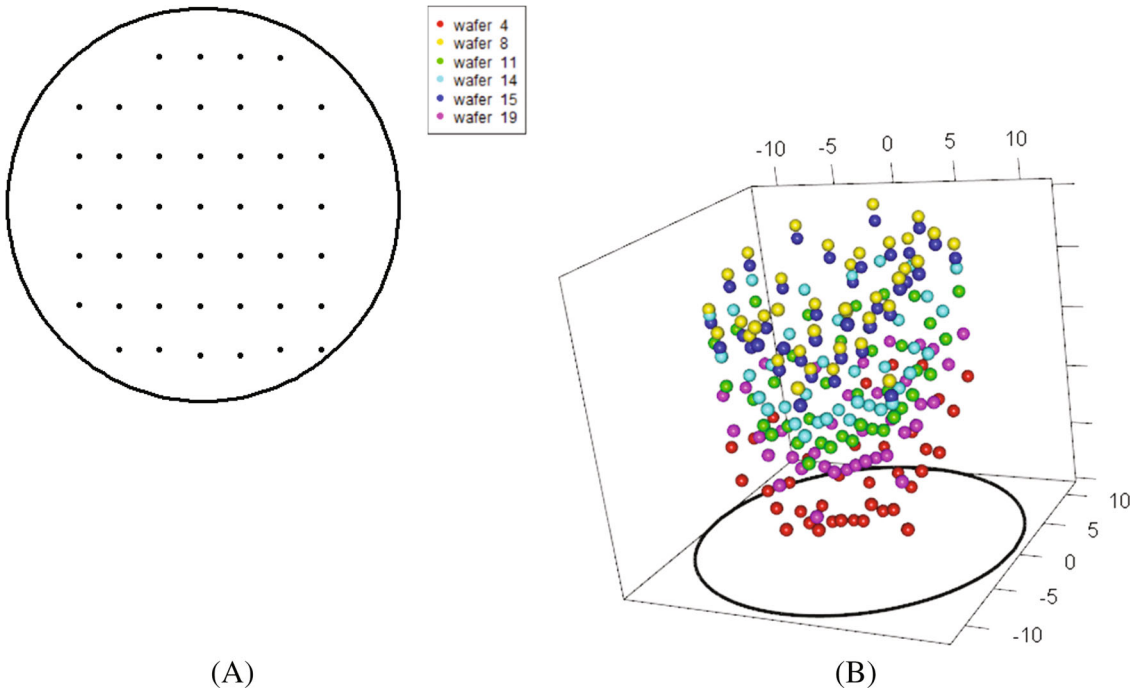

FIGURE 2 Sample grid (A). Wafer trench depth measurements in the considered lot (B)

Trench depth is fundamental for a proper functioning of the system; hence, its variability with respect to the product specification limits is carefully monitored and evaluated by appropriate indices.

In the dataset at hand, the trench depth is measured at 38 sample locations. The monitoring grid for a typical wafer of the lot is depicted in Figure 2A. The data refer to a batch of six wafers and the sampled depths are displayed simultaneously in Figure 2B where different colours identify measurements taken in different wafers. To display trench depths of different wafers together, depth measurements all refer to a typical wafer of the lot. The engineers of the company that provided the dataset found the process capability of the lot to be acceptable when the C_{pk} was as high as 1.4, slightly above the typical threshold of 1.33 usually adopted in capability analysis.

3 | CAPABILITY INDICES AND LOCAL CAPABILITY INDICES

Hereinafter, we consider the C_{pk} quantile approach introduced by Clements[3] that generalises the usual version of the index[2] to account for potential asymmetry of the distribution of the considered target variable, indicated now on by Y. In the application considered in this article, Y is the depth of trenches etched into a silicon wafer.

Clements's index is defined by the following equations:

$$C_{pu} = \frac{USL - \xi_{\tau_1}}{\xi_{\tau_2} - \xi_{\tau_1}} \quad C_{pl} = \frac{\xi_{\tau_1} - LSL}{\xi_{\tau_1} - \xi_{1-\tau_2}} \quad C_{pk} = \min\left(C_{pu}, C_{pl}\right) \tag{1}$$

where ξ_τ is the quantile of order $\tau, 0 < \tau < 1$ of Y and $\tau_1 < \tau_2$, typically $\tau_1 = 0.5$ and $\tau_2 = 0.99865$. LSL and USL represent the lower and upper specification limits of Y, respectively.

Clements's approach has been pivotal for the development of process capability index estimators for potentially non-normal data for a wide range of possible models of the target variable such as the Pearson and the Johnson system,[10,11] the Burr distributions,[12] the Weibull and lognormal distribution,[13] the t, gamma, and lognormal distributions,[14] and the zero-bound process,[7] just to mention a few. A detailed review is provided by Kotz and Lovelace.[4] To deal with non-normal processes, many researchers have also focused on methodologies based on transformation of the non-normal into normal data for the use of normal based PCIs. The Box-Cox transformation approach has the ability to produce good results, although it has not become very popular amongst practitioners because of the loss of the computed results with regard to the original scales.[15] Accurateness of PCIs for heavily skewed distributions has been discussed by Wu et al.,[16] whereas Kashif et al.[17] have demonstrated that C_{pk} values tend to be conservative of the true capability of the process for asymmetric distributions. In any case, relatively small C_{pk} values are, in general, associated with lower overall capability.

As mentioned in the introduction, the C_{pk} index is routinely computed at the lot level in the etching phase of the IC fabrication process; hence a unique C_{pk} value is calculated for the entire lot to assess the process ability to produce the output within the specification limits.

However, the assumption of a uniform capability of the process over the surface of a planar manufact cannot always be assumed or can be invalidated by empirical evidence. Local capability can be an issue, in particular, when the manufact has to be parcelled out to obtain individual items, as is the case for microchip production. In these circumstances, the homogeneity of the process capability can vary from piece to piece and deriving a local version of the C_{pk} may be worth it.

Despite the large use of the index in several contexts, almost nothing has been published concerning its use at a local level, an exception being the work of Borgoni and Zappa[6] for lognormal models.

In order to define a spatial version of C_{pk}, let $\{Y(\boldsymbol{s}), \boldsymbol{s} \in W\}$ with $W \subseteq \mathbb{R}^2$ be a random field indexed in the plane representing the characteristic of interest at any spatial location. In the case study presented below W represents the silicon wafer. Indicated by $\xi_\tau(\boldsymbol{s})$ the quantile of order τ of the conditional distribution of Y at location $\boldsymbol{s}$, a spatial version of the index in Equation (1) is given by

$$C_{pu}(\boldsymbol{s}) = \frac{USL - \xi_{0.5}(\boldsymbol{s})}{\xi_{0.99865}(\boldsymbol{s}) - \xi_{0.5}(\boldsymbol{s})}, C_{pl}(\boldsymbol{s}) = \frac{\xi_{0.5}(\boldsymbol{s}) - LSL}{\xi_{0.5}(\boldsymbol{s}) - \xi_{0.00135}(\boldsymbol{s})},$$
$$C_{pk}(\boldsymbol{s}) = \min\left(C_{pu}(\boldsymbol{s}), C_{pl}(\boldsymbol{s})\right). \tag{2}$$

A sample estimate of C_{pk} at each location $\boldsymbol{s}$ can be obtained by estimating the relevant quantiles $\xi_\tau(\boldsymbol{s})$ of Y at $\boldsymbol{s}$ and plugging these values, indicated by $\widehat{\xi}_\tau(\boldsymbol{s})$ from now on, in Equation (2), that is,

$$\widehat{C}_{pu}(\boldsymbol{s}) = \frac{USL - \widehat{\xi}_{0.5}(\boldsymbol{s})}{\widehat{\xi}_{0.99865}(\boldsymbol{s}) - \widehat{\xi}_{0.5}(\boldsymbol{s})}, \widehat{C}_{pl}(\boldsymbol{s}) = \frac{\widehat{\xi}_{0.5}(\boldsymbol{s}) - LSL}{\widehat{\xi}_{0.5}(\boldsymbol{s}) - \widehat{\xi}_{0.00135}(\boldsymbol{s})},$$
$$\widehat{C}_{pk}(\boldsymbol{s}) = \min\left(\widehat{C}_{pu}(\boldsymbol{s}), \widehat{C}_{pl}(\boldsymbol{s})\right). \tag{3}$$

In order to graphically represent the estimated C_{pk} over the entire wafer area, it is necessary to discretise the $C_{pk}(\boldsymbol{s})$ surface using a fine grid of wafer locations, that is, to preliminarily identify a set of points $G = (\boldsymbol{u}_1, \ldots, \boldsymbol{u}_N)$, with $\boldsymbol{u}_j \in W$ for $j = 1 \ldots, N$, calculate Equation (3) at each point of G and, finally, provide a raster representation of the estimated $C_{pk}(\boldsymbol{s})$ surface via a map. Ideally each point of G (or each pixel of the map) could represent a die eventually obtained from the wafer. We refer to G as *prediction grid* in the rest of the article.

Clearly, the measurement effort to monitor the capability of the production process at even a moderate spatial resolution is huge and a modelling approach has to be adopted to estimate the index surface.

In the next section, we suggest estimating local C_{pk} values in a fully non-parametric manner using additive quantile regression models.

4 | SEMIPARAMETRIC QUANTILE REGRESSION FOR TRENCH DEPTH SPATIAL SMOOTHING

Quantile regression was developed as an extension of the linear regression model.[18] Let Y be the response variable of interest and $\tau \in (0,1)$. The quantile regression model specifies the relationship between the quantile, ξ_τ, of the conditional distribution of Y and a set of p explanatory variables, $\boldsymbol{x}$, possibly depending upon a set of unknown parameters $\boldsymbol{\beta}_\tau$:

$$Q_y(\tau|\boldsymbol{x}) = \xi_\tau\left(\boldsymbol{x}, \boldsymbol{\beta}_\tau\right).$$

Alternatively, the model can be specified assuming

$$Y = \xi_\tau\left(\boldsymbol{x}, \boldsymbol{\beta}_\tau\right) + \varepsilon_\tau$$

where ε_τ is the error term whose τth quantile conditional on $\boldsymbol{x}$ is zero.

Direct quantile estimation is generally achieved by considering the following alternative definition of a conditional quantile

$$\xi_\tau(\boldsymbol{x}) = \underset{\xi}{\operatorname{argmin}}\, E\left[\rho_\tau\left(y_i - \xi\right)|\boldsymbol{x}\right].$$

where $\rho_\tau(u) = u \cdot (\tau - I(u < 0))$ is the so-called check function and $I(A < 0)$ is the indicator function of the event A, taking value 1 if A is true and 0 otherwise. Hence, considering the sampling counterparts, one obtains

$$\widehat{\xi}_\tau(\boldsymbol{x}) = \underset{\xi}{\operatorname{argmin}} \sum_{i=1}^{n} \rho_\tau\left(y_i - \xi\left(\boldsymbol{x_i}\right)\right).$$

In the linear case, the model writes as follows

$$Q_y(\tau|\boldsymbol{x}) = \boldsymbol{x}^T\boldsymbol{\beta}(\tau)$$

and the unknown parameters are estimated by solving

$$\widehat{\boldsymbol{\beta}}_\tau = \underset{\boldsymbol{\beta}_\tau \in \mathbb{R}^p}{\operatorname{argmin}} \sum_{i=1}^{n} \rho_\tau\left(y_i - \boldsymbol{x}_i^T\boldsymbol{\beta}_\tau\right).$$

The minimization problem above can be formulated as a linear programming problem and efficiently solved via linear programming methods. This approach is thoroughly described by Koenker.[19]

Flexible nonparametric specifications for the systematic component have been introduced by Koenker et al.,[20] and Koenker and Mizera[21] extended this approach to surface estimation using triograms.

In this article, we suggest modelling the quantiles of the depth of the trenches etched into a silicon wafer as a nonlinear function of spatial location $\boldsymbol{s} = (s_1, s_2)$:

$$Q_y(\tau|\boldsymbol{s}) = \xi_\tau(\boldsymbol{s}).$$

In particular, we consider a bivariate thin plate spline[22] to approximate the conditional quantile spatial field and the model writes as follows

$$Q_y(\tau|\boldsymbol{s}) = \sum_{k=1}^{K} B_k(\boldsymbol{s})\beta_{\tau k} \tag{4}$$

where $B_1(\boldsymbol{s}), \ldots B_K(\boldsymbol{s})$ is a bivariate thin plate spline basis function known and fixed whereas $\beta_{\tau 1}, \ldots, \beta_{\tau K}$ are unknown coefficients to be estimated from the data.

The semiparametric quantile regression is considered in a Bayesian paradigm since this approach was found to produce, in our experience, more stable results when extreme quantiles are of interest.

The Bayesian approach to quantile regression is more recent[23] and semiparametric quantile modelling in this context has been discussed by several authors (see amongst others Yue and Rue,[24] and Waldmann et al.[25]).

Hereinafter, we adopt the approach recently suggested by Fasiolo et al.[8] This approach bases quantile regression on a smoothed version of the check function, called the ELF (extended log-f) loss, defined by

$$\overline{\rho}_\tau(y;\lambda,\sigma) = (\tau-1)\frac{y}{\sigma} + \lambda\ \log\left(1+e^{\frac{y}{\lambda\sigma}}\right)$$

where $\sigma > 0$ is a scale parameter and $\lambda > 0$ such that $\overline{\rho}_\tau(\mathrm{u}) \to \rho_\tau(u)$ if $\lambda \to 0^+$ which allows more accurate quantile estimates and the use of efficient computational methods for model fitting. The estimation procedure is empirical Bayesian in nature. A gaussian 0 mean improper prior is assumed for $\beta_{\tau 1}, \ldots, \beta_{\tau K}$ in Equation (4) and their estimation is carried out maximising the a posteriori. The precision matrix is assumed to be equal to $\sum_{l=1}^{m}\gamma_l S_l$ where S_l is a positive semidefinite matrix scaled by γ_l. Instead $\gamma_1, \ldots \gamma_m$ as well as σ and λ are fitted to the data.

Since the additive quantile regression is based on the ELF loss, rather than on a probabilistic model for the response, the lack of a likelihood function does not permit Bayesian inference based on Bayes' rule to update the corresponding prior. The belief updating framework[26] can be adopted to perform the Bayesian updating via the loss function rather than the likelihood. The details of the estimation procedure are discussed by Fasiolo et al.[8] and implemented in the **qgam** R library.[27]

Once $\beta_{\tau 1}, \ldots, \beta_{\tau K}$ have been estimated using the approach sketched above, it is possible to obtain the estimated quantile spatial field $\widehat{\xi}_\tau(\boldsymbol{s})$ by replacing these values with their estimates $\widehat{\beta}_{\tau 1}, \ldots, \widehat{\beta}_{\tau K}$ in Equation (4) and, in turn, obtain the estimate $\widehat{C}_{pk}(\boldsymbol{s})$ at any spatial location of interest.

5 | EXPLORING PROCESS CAPABILITY AT THE LOT LEVEL

As mentioned in the introduction, the capability of the etching process is typically assessed at the lot level. However, wafers in the same lot may or may not be homogeneous in terms of their local capability and different shapes of the C_{pk} surface can be expected. In these circumstances, it is not appropriate to gather the wafers of the same lot to perform a global analysis on all wafers of the lot. In addition, when more than one wafer is of interest, a strategy is necessary to combine wafer measurements and perform a local capability analysis at the lot level.

In the rest of this section, we introduce two procedures based on resampling to address these issues, namely assessing homogeneity of local capability to identify possible homogeneous subgroups of lot wafers and combining information of different wafers to produce a local capability analysis at the lot level.

5.1 | Assessing homogeneity of local capability within a production lot

Process capability is typically addressed at the lot level; hence, the capability indices are calculated considering the measurements collected in all the wafers processed simultaneously in the same lot. For this index to be meaningful, it is necessary to assume that the production process possesses the same local capacity across wafers. In practice, this must be verified.

Hereinafter, we describe an algorithm based on resampling to evaluate local homogeneity of a set of wafers. The procedure aims to identify whether the process has the capability at any given location of a monitored wafer not inferior to the other wafers of the same lot. If a large proportion of the locations of this wafer are found to have poor capability with respect to the other wafers of the same lot, then it is to be classified in a separate subgroup. The idea of the procedure is to randomly take a wafer w from the lot and estimate $\widehat{C}^{w}_{pk}(\boldsymbol{u})$ at each location $\boldsymbol{u}$ of a prediction grid prefixed on the wafer area using the data collected in w. Considering the remaining wafers of the lot, one measurement is randomly sampled from those collected at each point of the sampling grid, obtaining a new sample of depth measures. The C_{pk} surface is fitted to this new sample and the procedure is iterated several times. For each location $\boldsymbol{u}$ of the prediction grid, the fifth percentile, denoted by $\mathrm{C}^{*}_{pk,0.05}(\boldsymbol{u})$, of the simulated estimates of $C_{pk}(\boldsymbol{u})$ is computed. Finally, an indicator variable is defined taking value 1 if the C_{pk} estimate at $\boldsymbol{u}$ in the wafer w initially considered is below this percentile and 0 otherwise that is, $I(\boldsymbol{u}) = 1$ if $\widehat{C}^{w}_{pk}(\boldsymbol{u}) < \mathrm{C}^{*}_{pk,0.05}(\boldsymbol{u})$.

If the percentage of 1's is greater than a threshold believed appropriate for the considered fabrication process, say 20%, then wafer w is removed from the lot since the process has, on average a poorer local capability than the other wafers of the lot. The procedure is re-applied by sampling another wafer of the lot until all the wafers have been inspected. The algorithm is described more in detail below.

Let W be a typical wafer of the lot L and $M = card(L)$. Let $G = (\boldsymbol{u}_1, \ldots, \boldsymbol{u}_N)$, with $\boldsymbol{u}_j \in W$ for $j = 1 \ldots, N$, be the prediction grid, $S = (\boldsymbol{s}_1, \ldots, \boldsymbol{s}_n)$ be the sample grid and $\mathbf{Y} = Y_1, \ldots, Y_n$ be the data sample. Let B be the number of sampling replicates.

Procedure 1.

Input: $\boldsymbol{Y}$, G, B, S, L, τ
Output: subsets of the lot wafers

set $temp \coloneqq L$
while $card(temp) \neq 0$
  Sample wafer $w \in L$ and calculate $\widehat{C}^{w}_{pk}\left(\boldsymbol{u_j}\right)$ for each $\boldsymbol{u_j} \in \boldsymbol{G}$.
  Set $L' \coloneqq L - w$ and $M' = M - 1$
  for $b \in B$
    for each $s_i \in S$ extract randomly one value Y from the M' available in L' and obtain a new sample $\boldsymbol{Y^*_b} = Y^*_{b1}, \ldots, Y^*_{bn}$
    calculate $\widehat{C}^{*}_{pk,b}\left(\boldsymbol{u_j}\right)$ for each $\boldsymbol{u_j} \in \boldsymbol{G}$ using $\boldsymbol{Y^*_b}$
  end for
  for $\boldsymbol{u} \in G$
    calculate $\mathrm{C}^{*}_{pk,\tau}(\boldsymbol{u}) = quantile\left(\left(\widehat{C}^{*}_{pk,b}(\boldsymbol{u}), b = 1, \ldots, B\right), \tau\right)$
    set $I(\boldsymbol{u}) = 1$ *if* $\widehat{C}^{w}_{pk}\left(\boldsymbol{u_j}\right) < \mathrm{C}^{*}_{pk,0.05}(\boldsymbol{u})$ and $I(\boldsymbol{u}) = 0$ otherwise
  end for
  if $\boldsymbol{100 \times N^{-1}} \sum_{\boldsymbol{u} \in G} I(\boldsymbol{u}) > \boldsymbol{\delta}$ set $L \coloneqq L'$ and $M = M'$
  set temp $\coloneqq L - w$
end while

As mentioned above, ideally, the spatial locations included in the prediction grid can be interpreted as potential dice obtained from the wafer. A by-product of the procedure is then to identify all the dice of the wafer where the process is of poor capability as compared to the dice in the same positions located in the other wafers of the lot. This also allows us to display those dice via a map that points out the portion or portions of the wafer area where variability with respect of the specification limits is critical.

5.2 | Combining local information of different wafers for lot local capability analysis

Once a subgroup of homogeneous wafers has been identified, it is necessary to combine the measurements taken at different points of each single wafer of the lot to perform a local capability analysis at the lot level. We propose below a resampling procedure for this task. The data of different lots are pulled together in a unique dataset. Being K the number of wafers in the set, one of the available K measurements is selected at random for each location of the monitoring grid to generate a resampled replicate of the data. The C_{pk} surface is estimated as detailed in Sections 3 and 4 using this replicate and the procedure is repeated numerous times. This permits one to estimate the distribution of the estimated C_{pk} conditional to each spatial location. Hence, a $1 - \alpha$ probability interval can be constructed using the simulated percentiles of this distribution for a selected value α. If the interval is entirely below some conventional value ε, the point is marked as one where the process is poorly capable. Repeating this procedure for all the points of interest, for instance for a prediction grid representing the location of the dice obtained by a typical wafer of the lot, allows one to build a map highlighting those dice of the lot where the process is expected to be too variable for the considered specification limits. The procedure is described in detail below. The notation used in the pseudo-code has been introduced in the previous subsection.

Procedure 2.

Input: $\boldsymbol{Y}$, G, B, S, ε, α
Output: capability map

for $b \in B$
  for each $\boldsymbol{s}_i \in S$ extract randomly one value Y_i^* from the K available and obtain a new sample $\boldsymbol{Y}_{\boldsymbol{b}}^* = Y_{b1}^*, \ldots, Y_{bn}^*$
  calculate $\widehat{C}_{pk,b}^*(\boldsymbol{u})$ for each $\boldsymbol{u} \in \boldsymbol{G}$ using $\boldsymbol{Y}_{\boldsymbol{b}}^*$
end for
for $\boldsymbol{u} \in G$
  calculate $\mathrm{C}_{pk,\alpha/2}^*(\boldsymbol{u}) = quantile\left(\left(\widehat{C}_{pk,b}^*(\boldsymbol{u}), b = 1, \ldots, B\right), \alpha/2\right)$

$$\mathrm{C}_{pk,1-\alpha/2}^*(\boldsymbol{u}) = quantile\left(\left(\widehat{C}_{pk,b}^*(\boldsymbol{u}), b = 1, \ldots, B\right), 1-\alpha/2\right)$$

  set $I(\boldsymbol{u}) = 1$ *if* $\left[\mathrm{C}_{pk,\alpha/2}^*(\boldsymbol{u}),\ \mathrm{C}_{pk,1-\alpha/2}^*(\boldsymbol{u})\right] < \varepsilon$ and $I(\boldsymbol{u}) = 0$ otherwise
end for
draw the map of $I(\boldsymbol{u})$

6 | NUMERICAL EXPERIMENTS

In this section, some results of a large simulation study are reported to evaluate the performance of the method described in Section 4. The quantile additive model is used to estimate quantiles adopting a bivariate thin plate spline (TPS) transformation of the spatial coordinates in the predictor of the quantile regression (QR). This model will be indicated by QRTPS. Equation (2) is used to estimate $C_{pk}(\boldsymbol{s})$ at any desired location $\boldsymbol{s}$.

The data were generated according to different scenarios in order to assess the performance of the proposed methodology with respect to: (1) an optimal parametric benchmark model and different sample sizes, (2) robustness to model specification, (3) robustness to outliers, (4) robustness to sample grid configurations.

6.1 | Comparison with optimal parametric models and the impact of the sample size

In the first scenario, the log-normal distribution is used. The parameters of the log-normal distribution, say μ and σ, are spatialised as a function of the point coordinates that is, $\mu(\boldsymbol{s})$ and $\sigma(\boldsymbol{s})$. Then, the quantile conditional to $\boldsymbol{s}$ are worked out for any desired $\boldsymbol{s}$ and $C_{pk}(\boldsymbol{s})$ is calculated as in Equation (2). In particular, we set $\mu(s_1, s_2) = \beta_0 + \beta_1 s_2$ and $\sigma(s_1, s_2) = \alpha_0 + \alpha_1 s_1^2 + \alpha_2 s_2^2$.

The quadratic shape for the variance surface is considered here (as well as in the next sections) following the usual assumption in response surface theory.[2] In our experience, modelling more irregular variance surfaces is definitely a challenging task to tackle with, in particular when the sample size is relatively small as in the case study considered in the article.

The parameters of the two equations are preliminarily estimated using the dataset at hand. This allows us to adopt the LSL and USL of the actual etching process that produced the data described in Section 2 when we calculate $C_{pk}(\boldsymbol{s})$.

The proposed approach is compared to the one suggested by Borgoni and Zappa[6] based on Generalised Additive Models for Location, Scale and Shape (GAMLSS,[28]) for log-normal data. Once the parameters of the log-normal distribution have been estimated using GAMLSS, the quantiles of interest are worked out to calculate $C_{pk}(\boldsymbol{s})$ at any desired location $\boldsymbol{s}$. The parameters of the log normal distribution conditioned to each spatial location were parametrized using the same equations reported above to generate the data. This model will be indicated by PCI-GAMLSS from now on.

In the simulation exercise, two different cases are considered assuming a sample grid of 45 and 52 points dislocated in a circular domain centred at the origin representing the wafer. These sample grids are superimposed on the maps in Figure 3A,B, where the C_{pk} surface, calculated as described above, is displayed.

Each Monte Carlo experiment is based on 1000 simulations. Figure 3C,D show the average $\widehat{C}_{pk}(\boldsymbol{s})$ surfaces in the two cases using QRTPS, whereas Figure 3E,F show the same values obtained using PCI-GAMLSS. Each surface is discretized

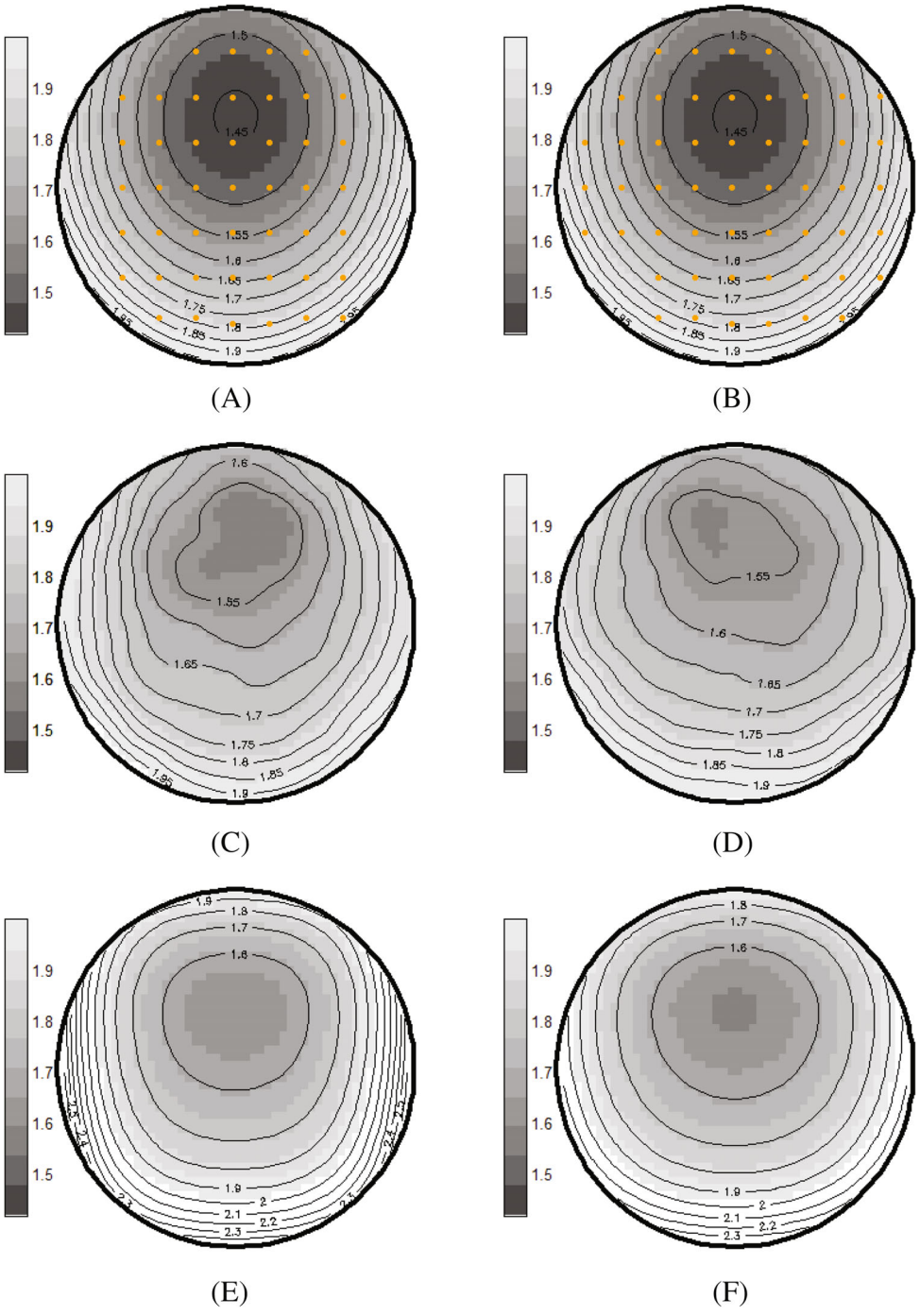

FIGURE 3 $C_{pk}(\boldsymbol{s})$ surface and sample grids: (A) and (B); averaged Monte Carlo QRTPS $C_{pk}(\boldsymbol{s})$ estimates: (C) and (D); averaged Monte Carlo PCI-GAMLSS $C_{pk}(\boldsymbol{s})$ estimates: (E) and (F). The colour scale of all plots of the panel is set using the deciles of $C_{pk}(\boldsymbol{s})$ values at the prediction grid depicted in figures (A) and (B)

by estimating $C_{pk}(\boldsymbol{s})$ over a grid of 1876 points. For each point, $\widehat{C}_{pk}(\boldsymbol{s})$ is computed using the two methods and the 1000 estimates are averaged to obtain a single value used to colour the corresponding pixel of the rasterized surface in Figure 3.

Figure 3 suggests that both approaches succeed in retrieving the shape of the C_{pk} surface picking up the correct pseudo paraboloid structure of the spatial index. However, the estimates obtained using PCI-GAMLSS appear to be too optimistic, estimating upwards the actual values of local C_{pk}, in particular in the south part and close to the border of the wafer, whereas the approach based on additive quantile regression produces estimates on average closer to the true value of the index.

6.2 | Robustness to model misspecification

In the second scenario, the robustness of the proposed approach to the stochastic model assumed to generate the data is considered. This time, the data are generated according to a symmetric distribution and, in particular, a generalised t-distribution is used. More specifically, at each sample location $\boldsymbol{s}$. the value of the target variable is generated setting $Y(\boldsymbol{s}) = \mu(\boldsymbol{s}) + \sigma(\boldsymbol{s})T$, where T is a central t-student random variable with 9 degrees of freedom. We assumed $\mu(\boldsymbol{s}) = \beta_0 + \beta_1 s_1 + \beta_2 s_2 + \beta_3 s_1^2 + \beta_4 s_2^2$ and $\sigma(\boldsymbol{s}) = \alpha_0 + \alpha_1 s_1^2 + \alpha_2 s_2^2$ in the simulation and, as above, the parameter of the two equations were preliminarily estimated using the dataset at hand. Figure 4A shows the true C_{pk} surface; Figure 4B,C show the average $\widehat{C}_{pk}(\boldsymbol{s})$ surfaces estimated by QRTPS using 45 and 52 points grids, respectively. Also, in this scenario, the C_{pk} surface is adequately reconstructed by the additive quantile model. The surface appears to be slightly overestimated particularly in case of the smaller sample grid.

6.3 | Robustness to outliers

In the third scenario, we considered the robustness of the procedure to outliers. The data are generated using the same constellation of parameters adopted in the first scenario but an increasing percentage δ of outliers is forced into the dataset, namely $\delta = 5\%, 10\%$, and 15%. These outlying observations were obtained by replacing $\delta\%$ of the values generated in scenario 1, with measurements drawn randomly from a uniform distribution in the interval $(0, y_\delta)$ where y_δ is the sample δ percentile of the simulated data. The panel in Figure 5 shows the results for the three considered cases using the sample grid of size 45, whereas the true C_{pk} surface and the sample grid are displayed in Figure 3A. It is found that, on average, the proposed procedure is quite robust to the presence of outliers. Comparing the maps in Figure 5 with Figure 3C (i.e., the estimates obtained when no anomalous data are present), we noticed that the C_{pk} values are only slightly increased in the presence of outlying observations and overestimation tends to mildly increase as the percentages of outliers gets larger.

6.4 | Robustness to sample grid configurations

Although a detailed discussion of the impact of the spatial shape of the sampling grid on the wafer area is beyond the scope of the present paper and has been discussed elsewhere (see for instance References 9,29), in the fourth scenario we briefly evaluated whether and how different spatial configurations of the spatial grid can impact the estimation of the C_{pk} surface. The data are generated using the log-Gaussian model with the same set of parameters adopted in the first scenario. In particular, the panel in Figure 6 shows the results for two different spatial configurations of a 45-points sample grid. In the first case, a sample grid allocated on concentric circles is considered. The points are allocated in space according to the optimal criterium discussed by Borgoni and Zappa.[29] In the second case, a complete spatial random sample is adopted where the 45 measurement points are drawn from a uniform distribution over the wafer region. Figure 6A,B show the sample points along with the actual C_{pk} surface, whereas Figure 6C,D display the estimated C_{pk} surfaces. In both cases, a reasonably good approximation of the C_{pk} surface is obtained on average, although it was found that the spatial allocation of points does impact estimation to some extent, a result that was somehow expected and found in other studies (see Reference 29). We just mention that using a log-normal GAMLSS model in this case provided a C_{pk} estimated map (not reported here) that nicely reproduced the true spatial shape shown in Figure 6A,B. However, akin to scenario one, C_{pk} values are remarkably estimated upward.

7 | CASE STUDY: ETCHING PROCESS CAPABILITY

7.1 | C_{pk} surface for trench depth

In this section, the proposed approach is applied to the wafer lot described in Section 2 where the trench depth is monitored by 38 sample locations displayed in Figure 2A.

The additive quantile model is applied to smooth the quantile surfaces of order 0.00135, 0.5, and 0.99865 for each wafer of the lot and Equation (2) is used to estimate the $C_{pk}(\boldsymbol{s})$ surface using the additive quantile model to estimate the quantile of interest at each location of the prediction grid. Results are reported in Figure 7 where the estimated $C_{pk}(\boldsymbol{s})$

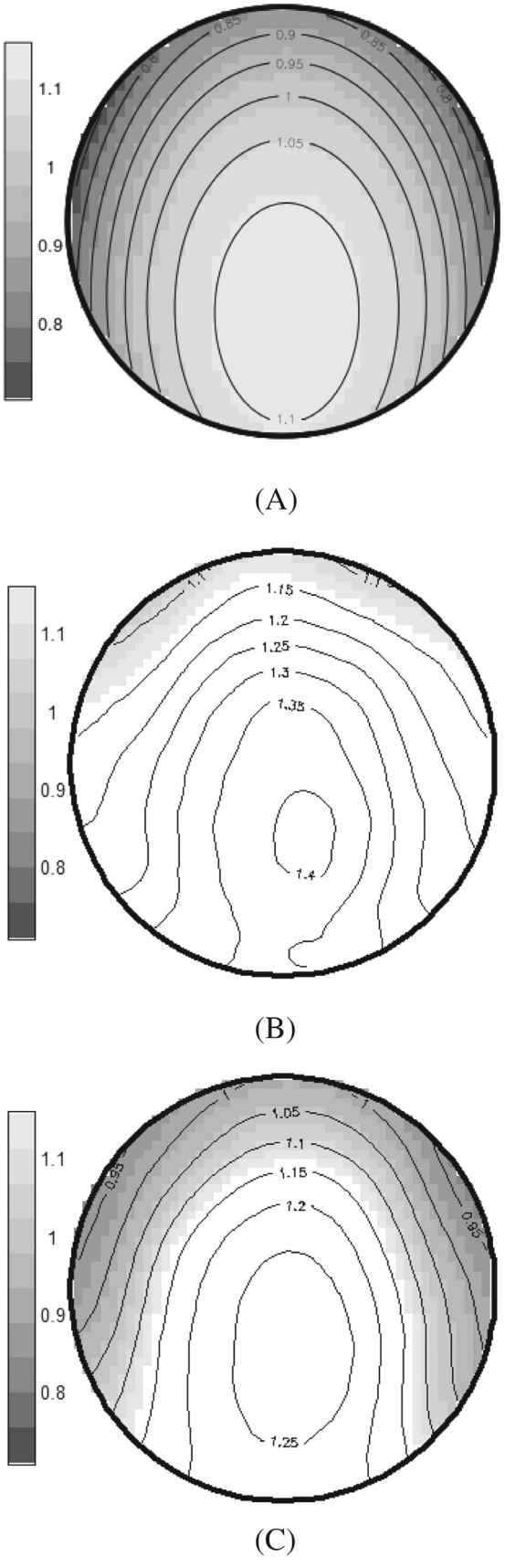

FIGURE 4 $C_{pk}(\boldsymbol{s})$ surface: true (A), averaged Monte Carlo QRTPS $C_{pk}(\boldsymbol{s})$ using a sample grid of size 45 (B) and using a sample grid of size 52 (C). The colour scale of all plots of the panel is set using the deciles of $C_{pk}(\boldsymbol{s})$ values at the prediction grid depicted in figure (A)

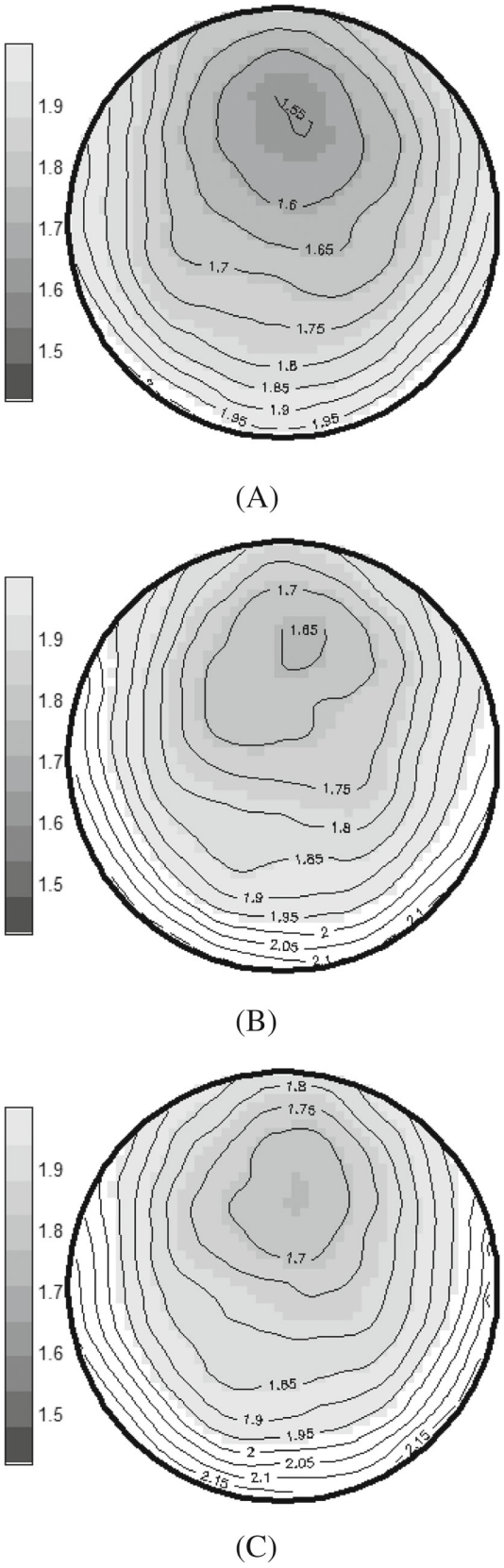

FIGURE 5 Averaged Monte Carlo QRTPS $C_{pk}(\boldsymbol{s})$ estimates in case of %5, (A), 10%, (B), and 15%, (C), of outlying observations. The colour scale of all plots of the panel is the same adopted in Figure 3A that is, using the deciles of $C_{pk}(\boldsymbol{s})$ values at the prediction grid depicted in figure (3A), which are clearly not affected by outliers

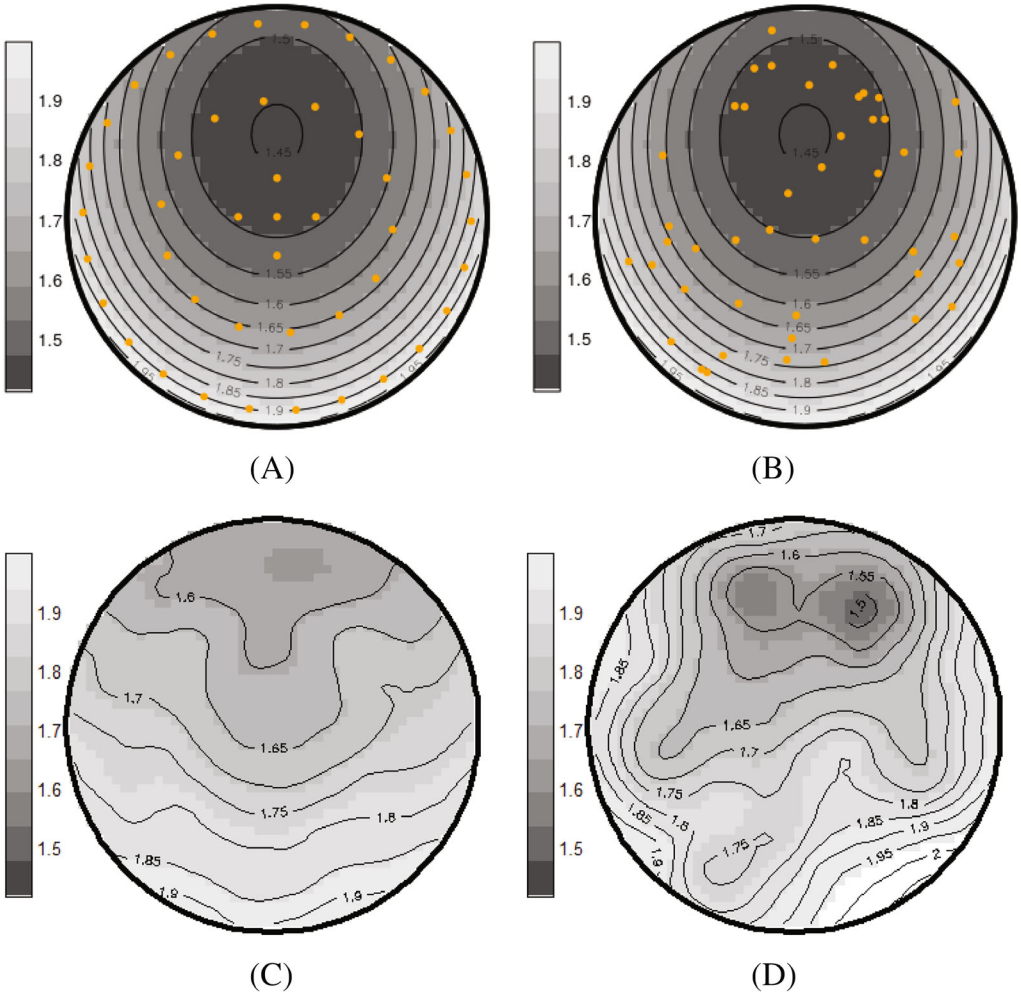

FIGURE 6 $C_{pk}(\boldsymbol{s})$ surface and sample grids with concentric circular and random points configuration: (A) and (B); averaged Monte Carlo QRTPS $C_{pk}(\boldsymbol{s})$ estimates: (C) and (D)

surface of each wafer of the lot is discretised using a grid of 1876 points internal to the wafer area. Two of the six wafers, labelled 8 and 15 in Figure 7, present substantial differences as far as the shape of the surface and the values of the index are concerned, the latter being much smaller than in the other wafers of the same lot.

The above analysis suggests, first, that the capability of the production process is not homogeneous over the wafer area, hence supporting the idea that the variability with respect to the specification limits of the etching process has to be evaluated locally that is, at each relevant spatial location. Second, the local capacity of the etching process may vary from wafer to wafer, hence making a standard lot-level capability analysis questionable.

7.2 | Etching capability analysis for lots

Although the capability of the process is typically assessed at the lot level, the previous analysis suggests that the process may be inhomogeneous across wafers. Hence, the local homogeneity of the process capability has to be preliminary explored. To this end, we applied Procedure 1 to identify potential subgroups of lot wafers where the etching process may suffer a too high variability with respect to the specification limits of the trench depth. The algorithm identified a remarkably poor local variability in wafers 8 and 15 since a large proportion, higher than one-fourth, of the dice has an estimated C_{pk} below the threshold denoted by $C^*_{pk,0.05}(\boldsymbol{u})$ in Procedure 1. In particular, 23.45% of the dice of wafer 8 have an estimated C_{pk} below the fifth percentile of the die-conditional distribution obtained via resampling whereas, in wafer 15, this percentage was found as large as 85%.

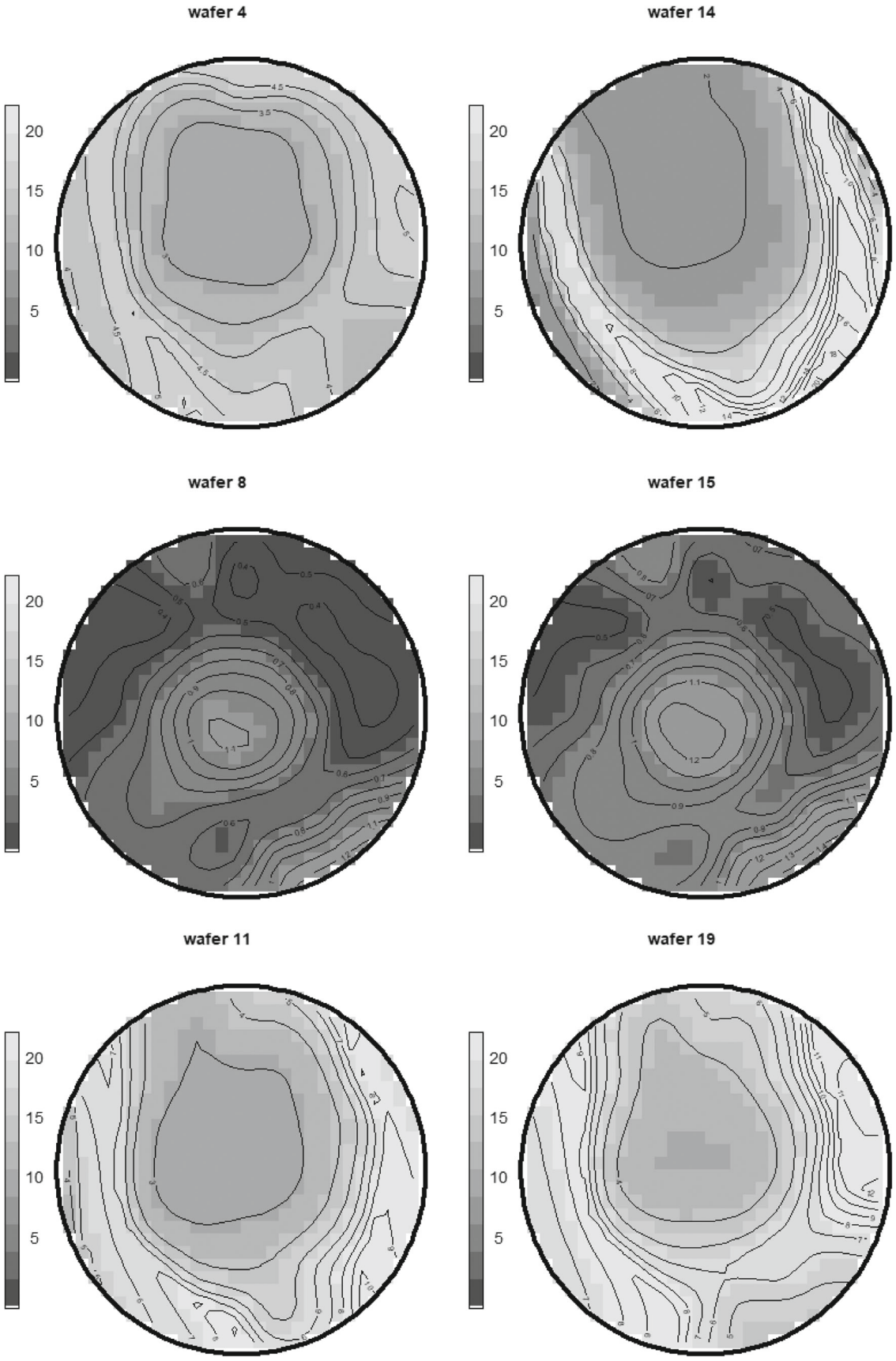

FIGURE 7 Estimated $\widehat{C}_{pk}$ surface for the six wafers of the lot. The colour scale is set by pulling together the C_{pk} estimates obtained at the points of the prediction grid of each wafer and calculating the decile of the overall set of estimates

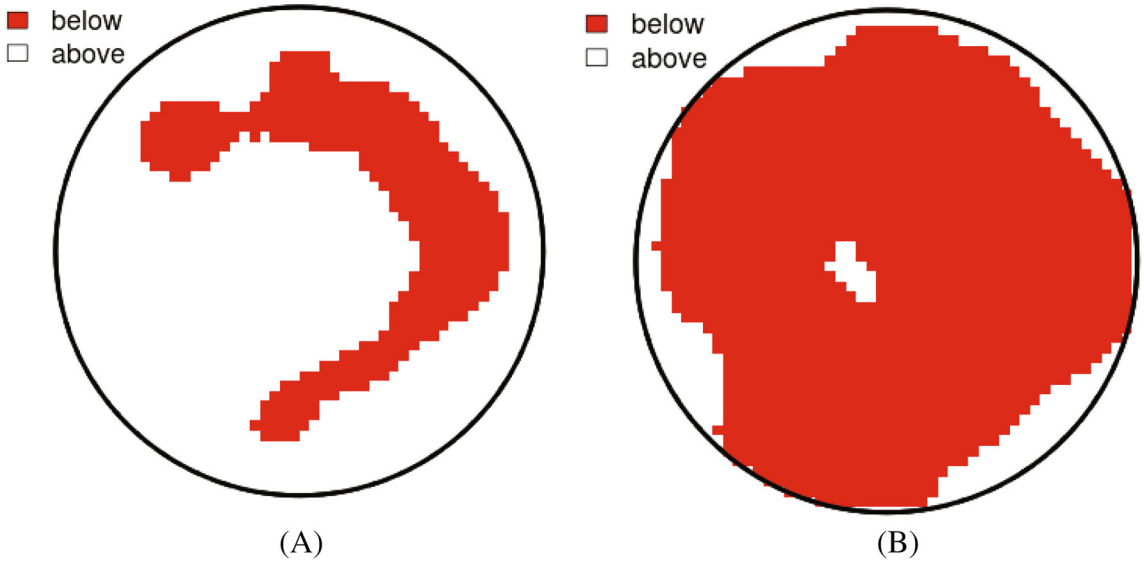

FIGURE 8 Dice with low local process capability: wafer 8 (A) and wafer 15 (B)

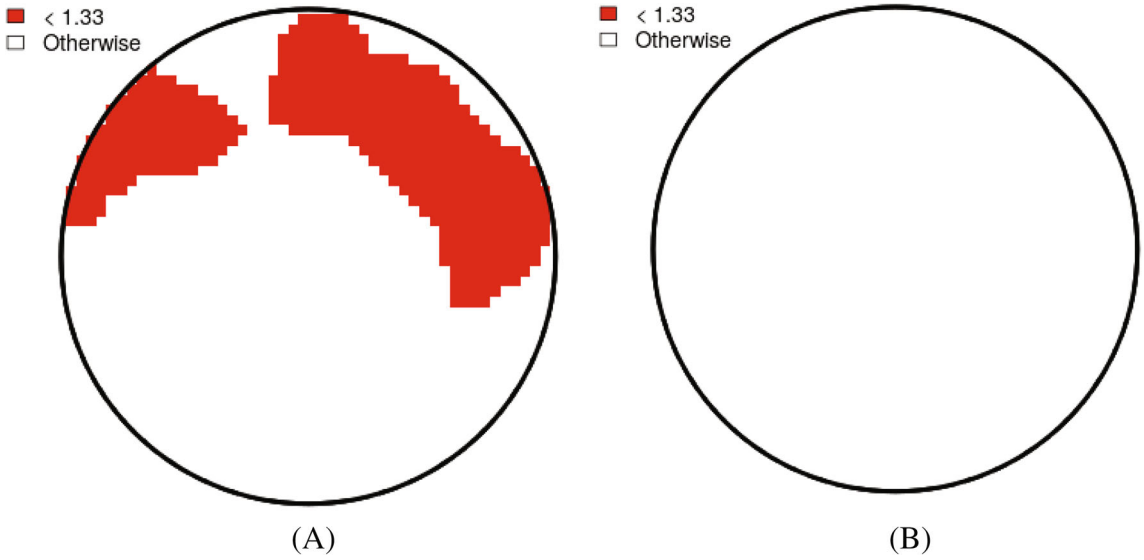

FIGURE 9 Maps identify in red the dice where the process has a low local capacity: subgroup 1 (A) subgroup 2 (B)

This algorithm, in addition, allows us to build a map representing those pixels of the wafer where low C_{pk} values occurred. The maps in Figure 8 display the result. Wafer 8 was the second wafer of the lot inspected by Procedure 2 whereas wafer 15 was the fifth. Similar results were obtained changing the order in which the wafers were evaluated.

The algorithm identified two sub-groups in the lot. The first includes the two wafers labelled 8 and 15 and the second includes all the other wafers of the lot.

In order to evaluate the local process capability for each of the two subgroups, Procedure 2 was applied using 1.4 as the threshold, that is, the usual target adopted in capability analysis. Clearly, other values more appropriate for specific situations can be adopted. The results are reported in Figure 9. Considering a *typical* wafer of the lot, the two maps highlight in red those dice of the wafer area where the probability interval obtained by algorithm 2 is below the considered threshold for each sub-group. For those dice, the process has a capability that can be considered significantly poor. In particular, in the first subgroup (wafer labelled 8 and 15), the percentage of the dice where the process is not capable is about 27.5% of the entire die yield (Figure 9A) whereas in the second sub-group the process capability was found appropriate for all dice; hence the entire surface of the wafer is coloured in white.

8 | CONCLUSIONS

PCIs are routinely used in fab practice to estimate the mean-variability performance of industrial processes with respect to both targets and specification limits. C_{pk}, in particular, is also commonly adopted for commercial ends. When the target variable is defined over a planar surface of a manufact, it is also relevant to assess the capability of the production process at any spatial location of the surface, in particular for those manufacts that have to be parcelled out to obtain single items.

Moving from the quantile specification of the C_{pk} introduced by Clements,[3] we proposed a methodology based on additive quantile regression to estimate the C_{pk} surface of the outcome variable on planar manufacts. We used bivariate thin plate splines to account for spatial regularities of the outcome and adopted a Bayesian framework for inference. Using simulation experiments, we demonstrated that the proposed approach manages to identify the actual C_{pk} surface in different scenarios and found it robust to model specification as well as to outlying observations.

We note that the sample size is also a relevant issue since grids of different size or spatial density may change the resolution of the C_{pk} surfaces and impact on the estimated variability. In microchip fabrication the number and the position of the sample locations are typically fixed at the beginning of the production, according to specific fabrication issues. The reduction of the sample size is only considered when the process moves from an experimental phase to production. A detailed investigation of the impact of modifying the shape or the size/density of the monitoring grid has been investigated elsewhere (Borgoni et al.[9] and Borgoni and Zappa[29]). However, this issue is beyond of the scope of the present article and has been only marginally considered in the simulation study.

We also suggested a procedure to evaluate whether the process capability is homogeneous amongst the wafers processed in the same lot when considered at the local level. This procedure excludes each wafer of the lot in turn and compares the C_{pk} surface estimated for this wafer with the others using an algorithm based on resampling. If the considered wafer has a low capability for a large portion (dice) of its area, it is removed from the lot and the procedure proceeds by inspecting another wafer. This approach allows the identification of those wafers of the lot characterised by poor capability. Finally, we proposed a second procedure to evaluate the local process capability for a group of wafers possibly identified by the previous algorithm. Using resampling, a probability interval of the C_{pk} is calculated conditioned to each spatial location. The interval is compared to a reference value and the process is considered having a low capability at the die level if the interval is below the considered threshold. In this way, we are able to draw a map that displays the portion of the wafer (dice) where the process has too high a variability with respect to the considered specification limits. The effectiveness of the two procedures has been exemplified using a dataset including georeferenced measures of the trench depth etched into a batch of wafers during the dry etching phase of the integrated circuits fabrication process.

ACKNOWLEDGMENT

We greatly acknowledge the DEMS Data Science Lab for supporting this work by providing computational resources. Open Access Funding provided by Universita degli Studi di Milano-Bicocca within the CRUI-CARE Agreement.

DATA AVAILABILITY STATEMENT

The data that support the findings of this study are available on request from the corresponding author. The data are not publicly available due to privacy or ethical restrictions.

ORCID

Riccardo Borgoni https://orcid.org/0000-0002-2520-3512
Diego Zappa https://orcid.org/0000-0003-4335-4530

REFERENCES

1. De-Felipe D, Benedito EA. Review of univariate and multivariate process capability indices. *Int J Adv Manuf Technol*. 2017;92:1687-1705.
2. Montgomery DC. *Introduction to Statistical Quality Control*. 6th ed. John Wiley & Sons, Inc; 2006.
3. Clements JA. Process capability calculations for non-Normal distributions. *Qual Prog*. 1989;22:95-100.
4. Kotz S, Lovelace CR. *Process Capability Indexes in Theory and Practice*. Arnold Publishing; 1998.
5. Hubele NF, Vännman K. The effect of pooled and un-pooled variance estimators on CPM when using subsamples. *J Qual Technol*. 2004;36(2):207-222.
6. Borgoni R, Zappa D. Model-based process capability indices: the dry-etching semiconductor case study. *Qual Reliab Eng Int*. 2020;36:2309-2321.
7. Lovelace CR, Swain JJ. Process capability analysis methodologies for zero-bound, non-normal process data. *Qual Eng*. 2009;21(2):190-202.

8. Fasiolo M, Wood S, Zaffran M, Nedellec R, Goude Y. Fast calibrated additive quantile regression. *J Am Stat Assoc*. 2021;116:1402-1412. doi:10.1080/01621459.2020.1725521
9. Borgoni R, Radaelli L, Tritto V, Zappa D. Optimal reduction of a spatial monitoring grid: proposals and applications in process control. *Comput Stat Data Anal*. 2013;58:407-419.
10. Bittanti S, Lovera M, Moirachi L. Application of non-normal process capability indices to semiconductor quality control. *IEEE Trans Semicond Manuf*. 1998;11:296-302.
11. Farnum NR. Using Johnson curves to describe non-normal process data. *Qual Eng*. 1996;9:329-336.
12. Castagliola P. Evaluation of non-normal process capability indices using Burr's distribution. *Qual Eng*. 1996;8:587-593.
13. Padgett WJ, Sengupta A. Performance of process capability indices for Weibull and lognormal distributions or autoregressive processes. *Int J Reliab Qual Saf Eng*. 1996;3:217-229.
14. Somerville SE, Montgomery DC. Process capability indices and non-normal distributions. *Qual Eng*. 1996;9:305-316.
15. Senvar O, Sennaroglu B. Comparing performances of Clements, box-cox, Johnson methods with weibull distributions for assessing process capability. *J Ind Eng Manage*. 2016;9:634-2016.
16. Wu CW, Chang CS, Pearn WL, Chen HC. Accuracy analysis of the percentile method for estimating non normal manufacturing quality. *Commun Stat Simul Comput*. 2007;36(3):657-696.
17. Kashif M, Aslam M, Al-Marshadi AH, Jun C. Capability indices for non-Normal distribution using Gini's mean difference as measure of variability. *IEEE Access*. 2016;4:7322-7330.
18. Koenker R, Bassett G. Regression quantiles. *Econometrica*. 1978;46:33-50.
19. Koenker R. *Quantile Regression*. Vol 38. Cambridge University Press; 2005.
20. Koenker R, Ng P, Portnoy S. Quantile smoothing splines. *Biometrika*. 1994;81:673-680.
21. Koenker R, Mizera I. Penalized triograms: total variation regularization for bivariate smoothing. *J R Stat Soc Ser B*. 2004;66:145-163.
22. Wood SN. *Generalized Additive Models: An Introduction with R*. Chapman & Hall/CRC Press; 2017.
23. Yu K, Moyeed R. Bayesian quantile regression. *Stat Probab Lett*. 2001;54(4):437-447.
24. Yue YR, Rue H. Bayesian inference for additive mixed quantile regression models. *Comput Stat Data Anal*. 2011;55(1):84-96.
25. Waldmann E, Kneib T, Yue YR, Lang S, Flexeder C. Bayesian semiparametric additive quantile regression. *Stat Model*. 2013;13(3):223-252.
26. Bissiri PG, Holmes CC, Walker SG. A general framework for updating belief distributions. *J Royal Stat Soc Ser B*. 2016;78(5):1103-1130.
27. Fasiolo M, Wood S, Zaffran M, Nedellec R, Goude Y. qgam: quantile non-parametric additive models; 2020. arXiv:2007.03303.
28. Rigby RA, Stasinopoulos DM, Heller GZ, De Bastiani F. *Distributions for Modelling Location, Scale and Shape: using GAMLSS in R*. CRC Press; 2019.
29. Borgoni R, Zappa D. Selecting subgrids from a spatial monitoring network: proposal and application in semiconductor manufacturing process. *Qual Reliab Eng Int*. 2017;33:1249-1261.

How to cite this article: Borgoni R, Farace VE, Zappa D. Non-parametric local capability indices for industrial planar manufacts: An application to the etching phase in the microelectronic industry. *Appl Stochastic Models Bus Ind*. 2022;38(5):884-900. doi: 10.1002/asmb.2673