PAN

Opto-Electronics Review

journal homepage: https://journals.pan.pl/opelre

A probabilistic approach for approximation of optical and opto-electronic properties of an opto-semiconductor wafer under consideration of measuring inaccuracy and model uncertainty

Stefan M. Stroka[1,2*], Christian Heumann[1], Fabian Suhrke[2], Kathrin Meindl[2]

[1] Department of Statistics, Faculty of Mathematics, Informatics and Statistics, LMU Munich, 80539 Munich, Germany
[2] ams-OSRAM International GmbH, 93055 Regensburg, Germany

Article info

Article history:
Received 24 Oct. 2022
Received in revised form 14 Mar. 2023
Accepted 30 Mar. 2023
Available on-line 12 May 2023

Keywords:
Gaussian process regression;
machine learning;
uncertainty quantification;
photoluminescence;
opto-semiconductor wafer measuring.

Abstract

This paper presents a probabilistic machine learning approach to approximate wavelength values for unmeasured positions on an opto-semiconductor wafer after epitaxy. Insufficient information about optical and opto-electronic properties may lead to undetected specification violations and, consequently, to yield loss or may cause product quality issues. Collection of information is restricted because physical measuring points are expensive and in practice samples are only drawn from 120 specific positions. The purpose of the study is to reduce the risk of uncertainties caused by sampling and measuring inaccuracy and provide reliable approximations. Therefore, a Gaussian process regression is proposed which can determine a point estimation considering measuring inaccuracy and further quantify estimation uncertainty. For evaluation, the proposed method is compared with radial basis function interpolation using wavelength measurement data of 6-inch InGaN wafers. Approximations of these models are evaluated with the root mean square error. Gaussian process regression with radial basis function kernel reaches a root mean square error of 0.814 nm averaged over all wafers. A slight improvement to 0.798 nm could be achieved by using a more complex kernel combination. However, this also leads to a seven times higher computational time. The method further provides probabilistic intervals based on means and dispersions for approximated positions.

1. Introduction

Nowadays, the Bayesian analysis is already being successfully applied in many research areas, like social sciences, ecology, genetics, medicine and more [1]. The particular characteristics of this probabilistic approach is that both observed and unobserved parameters receive a joint probability distribution. The Bayesian approach, which is based on the Bayes' theorem, is thus not only a model based on input data but also extends this with available knowledge about known model parameters [1]. After epitaxy of an opto-semiconductor wafer (wafer), important properties, like brightness or forward voltage, are measured by a destructive method, whereas other properties, for example the wavelength, can be measured non-destructively. The sampled measurements of the destructive process usually cannot completely and accurately reflect the properties of the entire wafer. To reduce the risk due to non-measured wafer positions, an approximation method is proposed. In the context of production, the given data fulfil the properties of spatial data. Many forms of approximations of spatial data have been used in the past. These algorithms are divided into deterministic methods, such as kernel approximation or spline interpolation, and stochastic methods, like spatial structure functions or radial basis functions (RBFs) [2]. Other frequently used algorithms are

*Corresponding author at: stefan.stroka@ams-osram.com

https://doi.org/10.24425/opelre.2023.145863
1896-3757/ Association of Polish Electrical Engineers (SEP) and Polish Academy of Sciences (PAS). Published by PAS

the local neighbourhood approach and the variational approach [3]. Although these established methods usually provide decent results, all methods are only able to provide approximations for certain wafer positions. However, production is affected by uncertainties due to measurement inaccuracy. A pointwise interpolation method does not consider variability in measuring. Thus, only measurement points are used instead of intervals considering input data uncertainty. Gaussian process regression (GPR) as a Bayesian approach can include uncertainties within the observed data and within the model itself through the joint probability distribution and, thus, constitute an approximation considering aleatoric and epistemic uncertainties.

1.1. Application-based GPR for approximation of a wafer

The basis of GPR is the selection of a prior mean and a covariance matrix or a covariance kernel function. This allows subjective knowledge about the wafer measurement to be used as prior information. At this point, it is also possible to consider input uncertainties as normally distributed errors within the prior. Based on the prior probability distribution and the likelihood, the algorithm results in a posterior probability distribution. This posterior follows a multivariate normal distribution which can be used to approximate values for unknown positions on the wafer. The prediction is made by weighting all possible predictions with the posterior distribution. The results are conditionally normally distributed predictions defined by their means and covariances [4].

1.2. Related research

In 2020, Barnes and Henn [5] compared machine learning (ML) algorithms such as RBF interpolation and GPR with a straightforward library lookup method for optical critical dimension (OCD) metrology. In this work, it is described that already 32 training points are sufficient for the ML method to be better than a library search. Schneider *et al*. [6] considered a Bayesian optimisation approach based on a GPR. Numerical simulation is used to reproduce measurement results of periodic micro- or nanostructures. These simulated structures are then described by optimised geometry parameters (geometry reconstruction). An earlier approach from 2015 by Henn *et al*. [7] attempts to obtain reliable estimates for quantitative characteristics of three-dimensional structures and associated realistic uncertainties by optimisable hybrid measurement techniques. A measurement method with a probabilistic prior and an approach with measurement methods combined through regressions are compared. Chen *et al*. [8] use an approach to increase the measurement accuracy of an optical scatterometry by using a fitting error interpolation-based library search method. A fitting error value is used to describe the wafer for a library search. Reference wafers are then those with the minimum difference in fitting error.

1.3. Aim of the paper

Destructive measurement methods can determine the opto-electronic properties of a wafer very well, but they are correspondingly cost-intensive and, therefore, increasing the number is not feasible. In practice, as well as in theory, methods for approximating or improving measurements in the field of opto-semiconductors are already being considered and applied. These approaches are mostly based on point estimators. The complex production of these wafers through the epitaxial process is difficult to control and measurement inaccuracies can also bias these values. A point estimator can deliver decent results compared to the test data but is not able to consider variability in measuring or systematic uncertainty. The aim of this work is, therefore, to get reliable approximations for the wafer measurements based on a probabilistic ML approach. The ML method focused on is the GPR, which is expected to provide robust point estimators and further quantifies uncertainty in the input data, as well as in the model. Comparisons are made with a state-of-the-art baseline method.

1.4. Paper organization

First, the necessity of new approaches in the context of production is shown in section 2. Second, the GPR and the baseline method are described in section 3. The experimental set-up is presented theoretically and practically in section 4. In section 5, results are presented and discussed, before a summary and an outlook for the future steps are given in section 6.

2. Front-end production optimisation with a Bayesian approach

Production at ams-OSRAM aims to manufacture high-quality opto-electronic semiconductors. Therefore, the front end of the production uses an epitaxy process to produce an epitaxial wafer from a substrate wafer as base carrier material. In the following sections, the paper focuses exclusively on a nitride-based process. This production process uses silicon carbide (SiC) as a base substrate or carrier material, respectively and grows gallium nitride (GaN)-based devices on it using metalorganic vapour phase epitaxy (MOVPE) [9]. According to Härle *et al*. [9], it is important for industrial production that, in addition to a stable epitaxy process, also a cost-effective chip technology is developed.

2.1. General idea and description

In the context of the opto-semiconductor process in a front-end production, different measurements are used to monitor the production step and achieve the best possible yield in subsequent further production steps. Measurement procedures are not part of the value chain. Hence, the goal of manufacturers is a maximum information gain with a minimum effort. The fundamental idea is to solve the problem of increasing the amount of information that is making predictions on unmeasured wafer positions, at the lowest possible additional cost using a state-of-the-art ML method and at the same time providing information about the reliability of the predictions based on a Bayesian approach.

2.2. Measuring systems

This paper considers the process steps after epitaxy and the subsequent measuring before further processing. In this regard, several tests are carried out to gain necessary information for further production steps. The main one is the so-called quick test (QT). The QT is a time-consuming and destructive procedure that provides information about the opto-electric properties. Since QT measured points are destroyed during the process, only a few points on the wafer (approximately 120) are tested. Increasing the number is, therefore, often not feasible. A non-destructive method is the photoluminescence (PL) measuring, which measures optical properties, like the wavelength. Here, information is obtained by irradiating the epitaxial surface of the wafer by photoexcitation [10]. PL measuring does not destroy the measured point but is also less accurate compared to the QT measuring.

3. Machine learning algorithms

In terms of application, a multivariate regression is needed to infer information from a higher dimensional space. These dimensions separate in our case into a spatial basis and associated measurements of a wafer. GPR is an approximation algorithm based on spatial dependencies of measurement points [11]. Therefore, the methodology of GPR and its probabilistic properties are described below. Furthermore, a multivariate approximation method based on RBFs is introduced and used as a comparison algorithm [12]. In general, the proposed algorithms can be applied to destructive and non-destructive measurement methods. To evaluate the analysis, the data set with (non-destructive) PL wavelength measurements is chosen.

3.1. Gaussian process regression

GPR is a non-parametric, probabilistic ML approach. The method is determined by Gaussian processes (GP) and uses Gaussian probability distributions. Instead of pointwise estimators, the probabilistic properties result in a distribution for the predictions. This allows to quantify uncertainties [13]. GP are stochastic processes with a finite set of random variables. Each random variable is a linear combination of normally distributed random variables and has, therefore, also a multivariate normal distribution. The paper applies the module Scikit-learn in Python [14], which is based on the presentation of Rasmussen [13]. The goal of GPR is to extract the information inherent in the observation without noise. For this purpose, the GP as multivariate normal distribution is used to model the observation without noise ε. The probabilistic GPR is defined by a posterior probability distribution. According to the Bayes' theorem, the posterior is determined by a prior distribution and the likelihood of actual observations Z. Let $Z = \{Z_i\}_{i\epsilon I} = \{Z_i = (x_i, y_i)\}_{i\epsilon I}$ be the observed data with x_i and y_i as the coordinates for the wafer position, I the associated finite index set and f the GP. Assuming that the observations without inherent noises can be represented by $f(Z)$, it follows:

$$\forall_{i\epsilon I}: w_i = f(Z_i) + \varepsilon_i = f((x_i, y_i)) + \varepsilon_i, \tag{1}$$

with w_i as the wavelength measurement. The prior distribution for each i in the sequence of random variables $\{w_i\}_{i\epsilon I}$ is normal since the error terms ε_i are normal and, therefore, defined by a mean and a variance. Thus, for the multivariate GP, the prior means are given by the expected value function $m_i = E(w_i) = f(Z_i)$ of the observations w_i, while the prior variance must be predefined as a covariance matrix, also called a kernel. For the application, two different kernels are considered with different levels of complexity which are introduced in section 4.1.

3.2. Radial basis function interpolation

RBF interpolation is a method for smoothing or multivariate interpolation of higher-order unstructured data [12]. The algorithm is based on RBFs or, equivalently, radially symmetric basis functions. According to Buhmann [12], a function is radially symmetric if the function value depends solely on the Euclidean distance from the origin. It follows that every function for which $\varphi(x) = \varphi(\|x\|)$ occurs is an RBF. Let Z be again the set of observations and f the inherent function without error ε_i. The aim of the RBF interpolation method is a continuous function s with the property

$$s(Z_i) = f(Z_i)\ \forall_{i\epsilon I} \tag{2}$$

which means that every training point Z_i as support point is met by the interpolation function s and s evaluated at Z_i equals the true value w_i without error ε_i for every Z_i. The algorithm defines the function s as a linear combination of basis functions. Let every φ_i be a basis function, which fulfils the condition of an RBF function, then

$$s(z) = \sum_{i\in I} \lambda_i \varphi_i(z), \tag{3}$$

with the scalar λ_i for every z within the interpolated value range. According to Fasshauer [15], this linear system can be solved uniquely only if the basis functions used are radially symmetrical. For the practical evaluation, the implementation of Scikit-learn [14] in Python is applied.

3.3. Uncertainty quantification

The key aspect of this paper is the quantification of uncertainty. While there are methods like 5-fold cross-validation that allow point estimation algorithms, such as RBF interpolation, to determine prediction uncertainty, these are not feasible in practical applications in the context of the experiment because of data sparsity. Measuring points are expensive and, therefore, only few data points per wafer are available. Hence, this paper focuses solely on the uncertainty quantification by GPR. Uncertainty quantification in a Bayesian approach is divided into two categories. The uncertainty within the data is called aleatoric uncertainty. In a physical approach, this is directly related to the measuring inaccuracy resulting from the measuring process. The second is the model uncertainty, also called systematic uncertainty. This quantifies the lack of knowledge, which is missing from the in theory perfect model [16].

3.3.1. *Confidence interval*

A confidence interval is defined by two bounds which are random variables and depend on a confidence level $\gamma = (1-\alpha)$ and a population of random samples. The confidence interval states that at a confidence level $\gamma \cdot 100\%$ the unknown parameter θ (e.g., the mean) is covered by the confidence interval at $\gamma \cdot 100\%$ for all repeatedly, randomly drawn samples of this distribution. In practice, the confidence interval depends only on a given population, meaning the training data. It can only quantify the aleatoric uncertainty.

3.3.2. *Prediction interval*

The prediction interval, like the confidence interval, is in this case a symmetric interval around the mean, defined by an upper and lower bound. Unlike the confidence interval, the limits of the prediction interval are determined based on the prediction error. The prediction interval uses the given information to describe which future observations of the same population are covered by the interval with a certain probability $\gamma = (1-\alpha)$ [13]. In the case of this paper, the prediction interval will be generated by sampling functions from the optimised GP. After fitting the posterior conditional distribution on the training data, it results in a family of not necessarily identical normal distributions equivalent to GP. From this family, an appropriate number of distribution functions are drawn as samples to describe which value ranges are covered by the interval to a fixed probability with the help of percentiles. The interval boundaries are defined by continuous functions, which also provide information about new observations of the same total population beyond the training data. This enables the quantification of uncertainties in both measurement and model accuracy.

4. Experimental setup

In the following, the ML-model setup and the practical application setup are presented before they are applied in section 5.

4.1. *ML model setup*

This section describes the structure of the application in a practical case and which fundamentals must be established for a reasonable implementation. For the practical part, there is a tuple of independent variables, the position data $Z = (x, y)$ on the wafer, and the dependent variable w as the measurement value. In practice, the only task necessary for the application of the GPR is the selection of a prior kernel. The prior represents the subjective view on the dependent variable and, therefore, cannot be unambiguously determined or at least not without very high additional effort. Consequently, two promising kernels were selected for this evaluation. The RBF kernel (squared exponential kernel) k_{RBF} as a standard kernel with an optimizable scalar λ and the length scale l

$$k_{RBF}(Z_i, Z_j) = \lambda * \exp\,(-\frac{d(Z_i, Z_j)^2}{2l^2}) \tag{4}$$

is considered at first for the simple model. A linear combination of squared exponential kernel, rational quadratic kernel, and maternal kernel with likewise optimizable scalars

$$\begin{aligned} k_c(Z_i, Z_j) = a \cdot \exp\left(-\frac{d^2}{2l^2}\right) + b \cdot \left(1 + \frac{d^2}{2\alpha l^2}\right)^{-\alpha} \\ + c \cdot \frac{1}{\Gamma(\nu) 2^{\nu-1}} \left(\frac{\sqrt{2\nu}}{l} d\right)^{\nu} K_\nu\left(\frac{\sqrt{2\nu}}{l} d\right) \end{aligned} \tag{5}$$

is considered next for the complex model. Thereby, the hyperparameters to be optimised are the length scale l, the smoothness ν, the scale mixture α and a, b, c as associated scalars. Further applies $d \coloneqq d(Z_i, Z_j)$ as a short form for the Euclidean distance, Γ as the gamma function, and K_ν as the modified Bessel function. The kernel combination can be generated by matrix addition and multiplication since each kernel satisfies the conditions as a covariance matrix [17]. Figure 1 shows the potential sample functions from GP for the respective prior distribution (kernel function). Comparing the GP sample function with a simple and complex kernel, a different degree of the GP functions variability can be seen. For more details regarding the kernels, it is recommended to compare with Rasmussen [13].

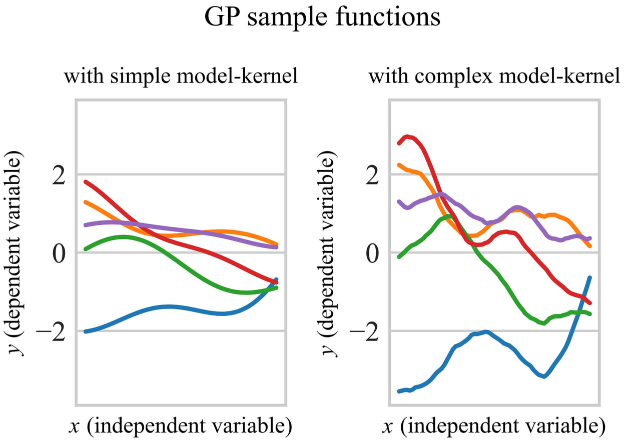

Fig. 1. Sampling y-values from GP with a given prior distribution (kernel). The GP is not yet trained and depends on mean and covariance function (kernel). Figure shows a two-dimensional prior distribution with one independent variable as an example, but more independent variables are also feasible.

4.2. *Practical application setup*

For the practical application, measurement data for a chip type in the blue colour range with nitride-based production processes were selected as an arbitrary prototype for the evaluation. The analysed wafer property in this paper is the wavelength. The QT measurement used in practice cannot be evaluated directly, as all up to 120 measurement points are necessary for training. For this reason, the PL measurement data set is used in the following showcase, as this is larger and thus test data are also given. To recreate the actual use case as realistically as possible, the PL data set is divided into test and training data. The wavelength values are measured for all given wafers of this chip type and used as the statistical population respectively as training and test data. For each wafer, an equidistant grid is used to declare approximately 120 measurement points of the data set per wafer as training data. This grid is identical to the pattern used for QT measuring. The remaining approximately 3000 measurement points are the test data. Due to the lack of several measured values for the same position, a measurement inaccuracy cannot be estimated directly. However, experts assume a certain uncertainty in the measurement of the wavelength with QT, which is confidential and may not be specified precisely. For transparency, an inaccuracy of 0.5 nm is used throughout the paper. This measurement error is constant and not wafer position-dependent, as the measurement of the edge point is carried out identically to the point within the inner area of the wafer. The used GPR implementation allows to define a specific measuring uncertainty as prior. Hereby, 0.5 nm will be added to the diagonal of the covariance matrix of the GPR. In the process of method application, each GPR is optimised individually for each wafer, resulting in a point estimate (mean vector) and a prediction interval (variance vector as the diagonal of a covariance matrix). The evaluation of the GPR with two different kernels is carried out in comparison to the described interpolation method as the baseline.

5. Results and discussion

Firstly, the results are evaluated based on a single, randomly selected wafer and uncertainties are quantified. Secondly, the evaluation is carried out empirically by considering the data of all wafers.

5.1. Results for an arbitrary wafer

Due to instability of the epitaxial growth, higher fluctuations occur in the edge region. In Table 1, a distinction is made between evaluation on the inner wafer area and evaluation on the complete wafer to represent the performance of the GPR more accurately. The inner wafer area is covered by the equidistant grid consisting of training data. Each measurement point in the complete test data set is part of the inner test data set if it lies within or on the perimeter line passing through the outer points of the equidistant grid. An insight into the results for one arbitrary wafer is given in Table 1.

Table 1.
Results of model fitting and prediction for an arbitrary wafer.

Method	LMLH[a]	RMSE (nm)		CT[b] (s)
		inner area	wafer	
RBF (baseline)	–	0.478	2.001	2.07
GPR (simple kernel)	−159.99	0.844	2.156	7.11
GPR (complex kernel)	−158.15	0.807	2.038	21.07

[a] log marginal likelihood
[b] computational time

Table 1 shows that based on the root mean square error (RMSE); the RBF interpolation model is performing best on both test data sets. The second-best model here is the GPR with a complex kernel. Comparing the two GPRs with different complexity, the high deviation in computational time (CT) is remarkable. Even though the GPR with the complex kernel performs better than the GPR with the simple kernel based on the RMSE and the optimised log marginal likelihood (LMLH) for both data sets, the CT almost triples. Comparing the ratio determined from RMSE divided by the wavelength median of the training data, it becomes clear that the proportional deviation and differences between them are small. The highest difference between proportional deviations is between the baseline and the GPR with simple kernel on the inner area test data set with 0.08%. It is noteworthy that exactly this difference in proportional deviation decreases when evaluating the complete test data set. The difference here is 0.034%. Thus, the GPR seems more stable than the baseline on the more difficult outer wafer area. Regarding the hypothesis, the RBF interpolation as the fastest method with the lowest RMSE can, therefore, dominate the comparison evaluations of point estimators. The disadvantage is obvious when considering uncertainty quantification, as the baseline interpolation method does not consider uncertainties. The lack of uncertainty quantification allows the method to compute the point estimator much faster than the GPR. Generally, GPR should still be preferred for practical purposes, since firstly, the deterioration in RMSE for point-wise regression is relatively small and secondly, a higher computational effort can be justified by a description of the model reliability. The following cross-sections from the wafer surface are used to obtain an insight into the regression and uncertainty quantification. These chosen sections are marked with red (vertical cross-section) and blue (horizontal cross-section) coloured lines within Fig. 2(b) and illustrated within Fig. 3. A cross-section considers the model and its results reduced by one dimension by setting one of the two independent variables x, y to zero. For the following graphical evaluation, the GPR model of the simple kernel is used for demonstration purposes. However, the same evaluation can be done with the complex kernel. Figure 2 shows all given historical data of this selected wafer in Fig. 2(a) and the uncertainty quantification with GPR for the same wafer in Fig. 2(b). The uncertainty is evaluated by using the standard deviation of the respective covariance matrix. A small standard deviation indicates a rather high certainty of the GPR model. Furthermore, those two certain cross-sections of the GPR will now be focused on. In addition to a point estimate, the GPR provides an uncertainty quantification in the form of a covariance matrix related to the wavelength as dependent variable, conditional on the position data. Figure 2(b), therefore, shows the top view of the model uncertainties of the GPR resulting from the regression on a selected wafer. This illustrates that the position dependence, respectively the direct distance to the closest training data point, is decisive for the reliability of the model. The position $(x, y) = (0, 1.1)$, in the outer area in Fig. 2 for example, shows that less training data in the area around the position results in a prediction with much higher prediction uncertainty. This can be seen in Fig. 2 by comparing the standard deviation value between the position $(x_1, y_1) = (-1, 0.59)$ for which measured values are available and the position $(x_2, y_2) = (1, 0.59)$, which is unknown during model training. In terms of practical application, this aspect is crucial, as measurement points

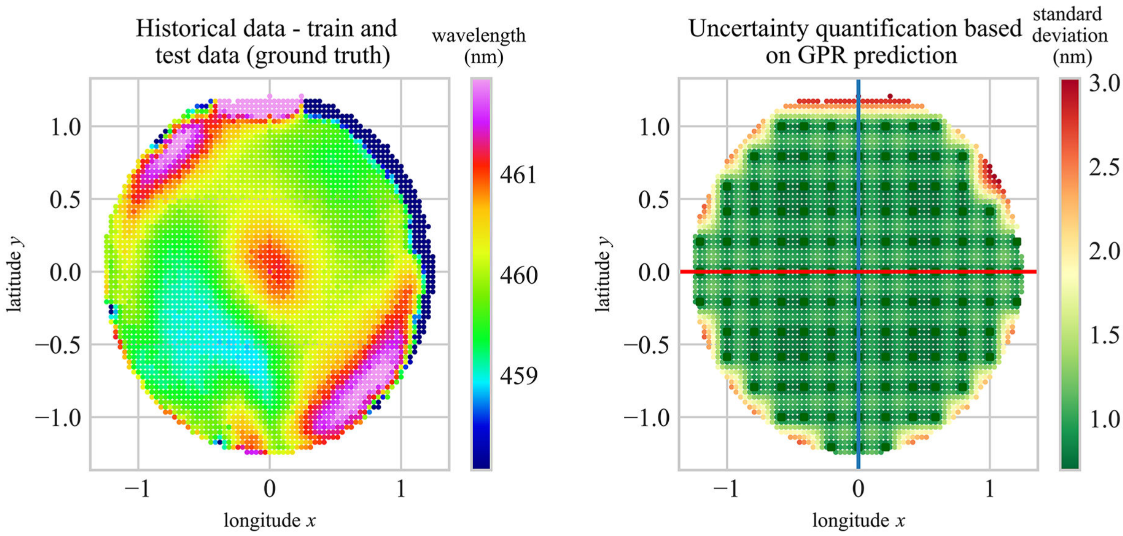

Fig. 2. Train and test data as ground truth (for this arbitrary wafer) (a) and position-based uncertainty quantified by GPR (b) in top view. Figure 2(a) displays the heterogenity of the wavelength for the given observations. GPR provides a standard deviation for every prediction, which indicates the certainty of the model for exactly this predicted position in Fig. 2(b). GPR is based on the PL train data (darkgreen dots).

may be missing in the production and the measurement process for undefined reasons. Information about the reliability of the model at these and surrounding measurement points is therefore essential. Now, the graphical observation is reduced to a section, for this purpose the longitude $x = 0$ in Fig. 3(a) and the latitude $y = 0$ in Fig. 3(b) are set to display the results in a side view. Both models are trained with the complete training data set (not only with data of each cross-section). Figure 3(a) and Fig. 3(b) show the point estimation and the confidence and prediction intervals resulting from the GPR. The graphical comparison of the RBF interpolation curve and the GPR prediction curve for the vertical cross-section in Fig. 3(a) shows that both are almost identical for most of the definition range. Deviation can only be observed in the boundary areas $[-1.21, -1.0]$ and $[1.0, 1.21]$. These deviations become more obvious when comparing these point estimators to the test data. Looking at the model uncertainties, the right border area $[1.0, 1.21]$ has high uncertainties, which results from a missing measurement point at $y = 1.21$. The model, therefore, extrapolates at this point. The same applies to the horizontal cross-section in Fig. 3(b). Both methods are often visually approximately congruent, yet both cannot completely reproduce the test data without deviation. The right border $[0.5, 1.21]$ is noticeable. Here, the RBF interpolation and the GPR prediction diverge strongly in some cases, and yet both fail to recognise the actual trend. This results from the fact that

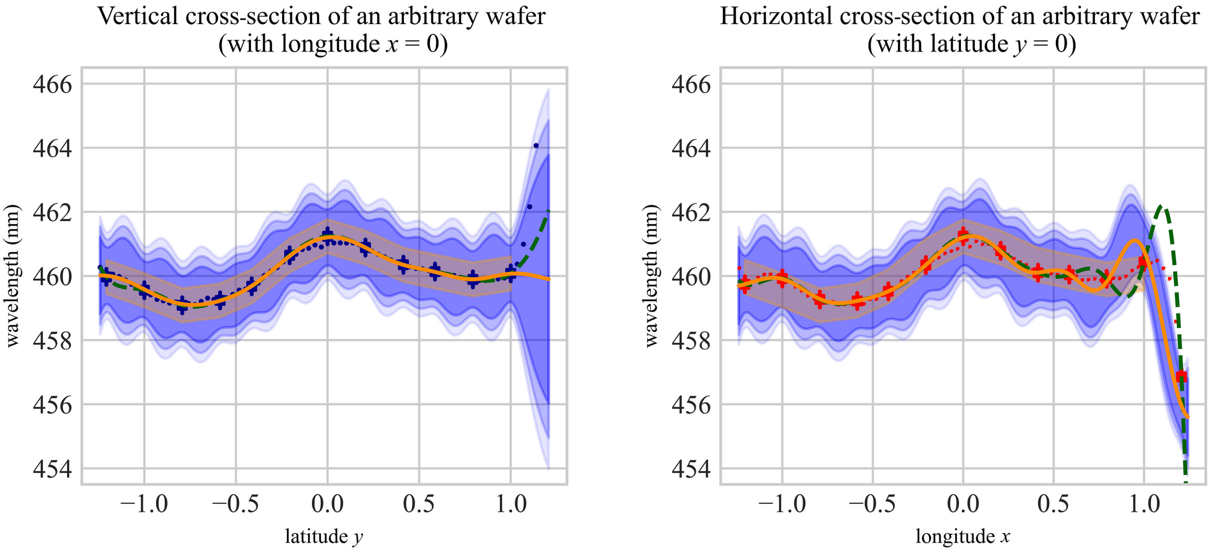

Fig. 3. Two different cross-sections of the same wafer. Both figures show the RBF interpolation (green line), the GPR (orange line), and the real observation, respectively the test data (left: darkblue/right: red dotted line). The results of the green and red line methods are similar for a large part of the range of values where the lines overlap. Each model is trained with the identical training data set. Squared points indicate the PL train data on the cross-section (left: darkblue/right: red squares). Horizontal lines through the training data indicate the fixed measuring inaccuracy (0.5 nm). The narrow interval across the value range (orange) shows the confidence interval for 97.5% probability, whereas the broader intervals (different shades of blue) mark the prediction interval for different probabilities (from the inside out: 90%, 95%, and 97.5%).

both models aim to consider the relatively low wavelength value at $x = 1.21$ from the training data. It is also noteworthy that in Fig. 3(b) in comparison to Fig. 3(a), there is another training point at the right-hand border, from which it follows that the model uncertainty in this area is significantly lower. The relevance of the prediction intervals for the application should be pointed out once again. By method application, there is no longer only one measured value in production that is close to reality only in the optimal case, but an interval range that covers reality with a fixed probability compared to the GPR model. Furthermore, it also becomes graphically clear that the difference between the baseline and the GPR is only small in relative terms and can be neglected regarding the added value due to the prediction interval.

5.2. *Limitation of the ML methods regarding the practical application*

In section 5.1, it is assumed, that approximately 120 measuring points of a wafer are used as model training points. In practice, measurement points can get destroyed during production or measuring. Figure 4 shows that each method needs a certain minimum number of training points to achieve good results.

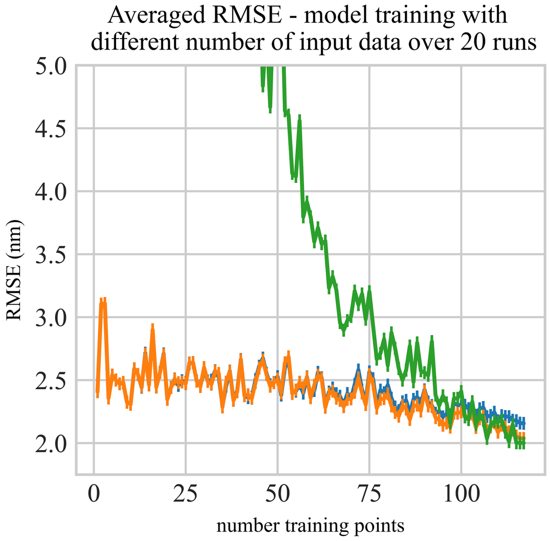

Fig. 4. Evaluation based on averaged RMSE with different numbers of input data. Each model is trained with a fixed number of random sampled training points (x-axis). To obtain robust results, 20 runs are performed for each number of random sampled training points. Sampling is done to simulate the loss of a training point. The evaluation is based on the complete wafer test data set with about 3000 points. The figure shows the RBF interpolation (green line), the GPR with simple kernel (blue line), and the GPR with complex kernel (orange line).

The evaluated models and kernels are trained with different numbers of input data. For this purpose, the respective number of training points is randomly drawn from the training data used in section 5.1. For each fixed number of training points, 20 identical models are trained with different samples and then the mean RMSE is computed. This is necessary because the random sampling of training points for each model has a strong influence on the prediction model. Figure 4 thus shows the tendency of all three approaches to worsen when the number of input data is reduced. In the range above 100 training points, there is hardly any decrease in RMSE value resulting from fewer points. With less than 100 training points, however, a strong deterioration of the RBF interpolation becomes apparent. This worsens up to a maximum RMSE of approx. 14 000 nm, for that reason it cannot be illustrated nicely in Fig. 4 and is therefore truncated. In the comparison of the two GPRs with different kernels, a slight tendency towards deterioration is recognisable in both. It is also noticeable that the more complex kernel cannot deliver an improvement compared to the simple kernel below the minimum number of points. Even though the RMSEs are averaged, the variability of the results increases for smaller numbers of training points. Although this is a limitation for the GPR, the RBF interpolation becomes significantly worse and unreliable at less than 100 training points and below. This is one major advantage of the GPR over the baseline, since in practical application not every single wafer can be checked on its own. GPR provides reliable results even if the number of wafer measurements is exceptionally below 100 points.

5.3. *Empirical results over all wafers*

Table 2 shows the lowest RMSE and subsequently the best overall point estimation achieved by the baseline model. Comparing the GPR models, the smaller LMLH value shows that the model with the complex kernel achieved a much better model fit on train data. However, this is not reflected by the mean of RMSE values related to the test data where both results are similar. When looking at the CT, GPR with the simple kernel takes about 53 times longer than RBF interpolation. The ratio is even higher for the GPR with complex kernel, where it needs about 366 times as much CT as the baseline. It also follows that using the complex kernel instead of the simple kernel takes approximately seven times more CT.

Table 2.
Results of model fitting and prediction for all wafers from the given population.

Method	**LMLH**[a]	**RMSE (nm)**		**CT**[b] **(s)**
		inner area	wafer	
RBF (baseline)	–	0.224	0.766	0.06
GPR (simple kernel)	−116.65	0.303	0.814	3.18
GPR (complex kernel)	−100.82	0.273	0.798	22.01

[a] log marginal likelihood
[b] computational time

Figure 5 shows the RMSE between measured values and predictions of the models for the inner and the complete wafer test data sets. In relation to the evaluation results of other wafers in the population, the results for the randomly selected wafer from Table 1 are in the upper outlier range.

The overall best performing algorithm in Fig. 5 is the RBF interpolation with the lowest RMSE median for both test data sets. The GPR with a simple kernel and the model with a complex kernel are worse than the baseline at the

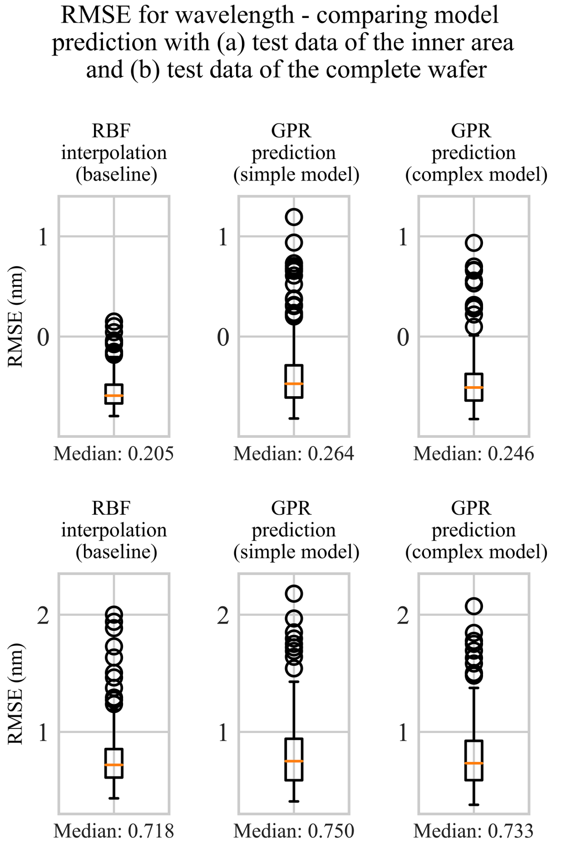

Fig. 5. Evaluation of the empirical results of all models with RMSE for the wavelength for 239 wafers. Each model is trained and tested using the measurements of a specific wafer of the inner area (a) or of the complete wafer (b).

median by 0.059 nm and 0.041 nm for the inner area test data set and 0.032 nm and 0.015 nm for the complete wafer test data. This is further confirmed by comparing the mean values from Table 2. It is noteworthy that the relative difference between the baseline and the GPR models regarding the point-wise prediction error decrease from the application on the inner area to the application on the complete wafer. This indicates a tendency, as in section 5.1, where the GPR is more stable on data with higher variability than the baseline. In summary, the RBF interpolation empirically performs better as a point-wise estimator than the GPR. Based on the results in Table 2, it is evident that the model fitting with a more complex kernel performs better on the training data but hardly represents an improvement compared to the actual test data, especially when considering the longer CT. This indicates an overfitting of the model and, consequently, a possibly unnecessarily high kernel complexity. As already stated in section 5.1, the differences in the mean and median are small in relative terms. These are also significantly lower than the measurement inaccuracy (aleatoric uncertainty). In terms of production, it can be considered not significant. The apparently very high CT compared to the baseline is not particularly noticeable in terms of a real-world application, considering that the regression of a wafer with a simple kernel only takes approximately 3 s. Nevertheless, it must be noted that this modelling was not done using powerful computers and yet a reasonable time was achieved considering the good results for the point estimator and the additional value for production through uncertainty quantification.

6. Conclusions and outlook

Highly complex industrial manufacturing relies on meaningful measurements in the production process to achieve the highest possible and most qualitative yield. Stable and reliable measurements are highly time-consuming and costly, which is why production must get along with sparse measurements and, therefore, accept uncertainties. The proposed probabilistic GPR provides point estimates considering uncertainties of measurement and model. The normally distributed GPR gives a continuous wafer map regarding the measured properties with equally continuous uncertainty quantification for the whole wafer. Although the analysis is carried out for wavelength, any other property can also be investigated in further research. The empirical evaluation for the wavelength shows that a GPR with an RBF kernel as simple kernel is sufficient to achieve an average RMSE of 0.303 nm on the inner area and 0.814 nm on the complete wafer. In relative comparison to the specified measurement accuracy of 5 nm, the fitting error is low and not significant in terms of production. The RBF interpolation as baseline method and the GPR with complex kernel surpass this result only barely with an RMSE of 0.224 nm (inner area) and 0.766 nm (complete wafer), and 0.273 nm (inner area), and 0.798 nm (complete wafer). A point estimator has often little significance when used productively, as it cannot always provide a reliable prediction. This fact poses a great risk to the goals of scrap minimisation, compliance with specification limits, and yield maximisation. A high variability due to the measuring process after epitaxy of a wafer increases this risk significantly. In detail, outliers that exceed the accepted variability limits equivalent to the prediction interval based on measurement and model uncertainties can highlight possible specification failures and thus serve as an alarm system. Measurement and model uncertainties are, therefore, important selection criteria in chip production to estimate dimension of the problem and amount of chips that will be out of specification and the associated yield loss in later production steps. Wafers are selected according to uncertainties for the best possible further processing or also for certain specifications. The GPR, unlike the baseline, has a probabilistic uncertainty quantification. Therefore, GPR directly enables a meaningful uncertainty-based classification of the output to meet the needs of the production. Against the clear advantages stands a higher CT. A GPR with a simple kernel takes on average 53 times longer than baseline interpolation for a complete wafer. Regression with a complex kernel takes even longer, at around 366 times the runtime of the baseline. In productive terms, however, GPR with a RBF kernel can determine a continuous point estimator with uncertainties of a whole wafer in about 3 s, which is why the computing times can be accepted in current applications. Even a GPR with a low-complexity kernel thus offers all the advantages necessary for production, both through exact point estimators and through the determination of uncertainties in measurement and model. Besides the focus on a safety-based categorisation of the output, it is equally important to extend the view to the overall production. The results from GPR are wavelength measurements or intervals of a whole wafer after epitaxy. These are only the results of an intermediate process step.

From an overall production perspective, the GPR results should further be used to optimise the subsequent process steps for chip production. Currently, the approximately 120 measuring points considered in the paper are used to conclude the resulting number and quality of the chips with the help of a regression approach. Viewed holistically, the GPR can thus be seen as a pre-processing step for this regression. Building on the probabilistic approach, a Bayesian regression model, such as a Bayesian neural network can also be used. Possibly, even the prediction intervals respectively to the inherent standard deviation of the GPR model could find further use as a meaningful prior distribution. Furthermore, the significantly higher number of input measurement points resulting from the model should also offer an improvement for any regression. Thus, it can be pointed out that the probabilistic GPR approach provides a solid improvement opportunity for the studied area and offers a clear potential concerning the further process optimisation.

References

[1] van de Schoot, R. *et al*. Bayesian statistics and modelling. *Nat. Rev. Methods Primers* **1**, 1–26 (2021). https://doi.org/10.1038/s43586-020-00001-2

[2] Myers, D. E. Spatial interpolation: an overview. *Geoderma* **62**, 17–28 (1994). https://doi.org/10.1016/0016-7061(94)90025-6

[3] Mitas, L. & Mitasova, H. Spatial Interpolation. in *Geographical information systems: principles, techniques, management and applications (*eds. Longley, P. A., Goodchild, M. F., Maguire, D. J. & Rhind, D. W.) 482–492 (Wiley, 1999).

[4] Rasmussen, C. E. Gaussian Processes in Machine Learning. in *Advanced Lectures on Machine Learning* (eds. Bousquet, O., von Luxburg, U. & Rätsch, G.) 63–71 (Springer, 2004). https://doi.org/10.1007/978-3-540-28650-9_4

[5] Barnes, B. M. & Henn, M.-A. Contrasting conventional and machine learning approaches to optical critical dimension measurements. *Proc. SPIE* **11325**, 222–234 (2020). https://doi.org/10.1117/12.2551504

[6] Schneider, P.-I., Hammerschmidt, M., Zschiedrich, L. & Burger, S. Using Gaussian process regression for efficient parameter reconstruction. *Proc. SPIE* **10959**, 200–207 (2019). https://doi.org/10.1117/12.2513268

[7] Henn, M.-A. *et al*. Optimizing hybrid metrology: rigorous implementation of Bayesian and combined regression. *Proc. SPIE* **14**, 044001 (2015). https://doi.org/10.1117/1.JMM.14.4.044001

[8] Chen, X., Liu, S., Zhang, C. & Zhu, J. Improved measurement accuracy in optical scatterometry using fitting error interpolation based library search. *Measurement* **46**, 2638–2646 (2013). https://doi.org/10.1016/j.measurement.2013.04.080

[9] Härle, V. *et al*. GaN-Based LEDs and Lasers on SiC. *Phys. Status Solidi A* **180**, 5–13 (2000). https://doi.org/10.1002/1521-396X(200007)180:1<5::AID-PSSA5>3.0.CO;2-I

[10] Stern, M. L. & Schellenberger, M. Fully convolutional networks for chip-wise defect detection employing photoluminescence images. *J. Intell. Manuf.* **32**, 113–126 (2021). https://doi.org/10.1007/s10845-020-01563-4

[11] Oliver, M. A. & Webster, R. Kriging: a method of interpolation for geographical information systems. *Int. J. Geogr. Inf. Syst.* **4**, 313–332 (1990). https://doi.org/10.1080/02693799008941549

[12] Buhmann, M. D. *Radial Basis Functions: Theory and Implementations*. (Cambridge University Press, 2003).

[13] Pedregosa, F. *et al*. Scikit-learn: Machine learning in Python. *J. Mach. Learn. Res.* **12**, 2825–2830 (2011). https://jmlr.org/papers/volume12/pedregosa11a/pedregosa11a.pdf

[14] Fasshauer, G. E. *Meshfree Approximation Methods with MATLAB*. (World Scientific, 2007).

[15] Walker, W. E. *et al*. Defining uncertainty: a conceptual basis for uncertainty management in model-based decision support. *Integr. Assess.* **4**, 5–17 (2003). https://doi.org/10.1076/iaij.4.1.5.16466

[16] Patel, J. K. Prediction intervals – a review. *Commun. Stat. – Theory. Methods* **18**, 2393–2465 (1989). https://doi.org/10.1080/03610928908830043

[17] Duvenaud, D. *Automatic Model Construction with Gaussian Processes*. (University of Cambridge, 2014). https://doi.org/10.17863/CAM.14087