Implementation of Spatio-Temporal Wafer-Level Correlation Modeling with Progressive Sampling and Wafer-based clustering

Abhishek M Patil – amp170830, Shri Prakash – sxp175131

Department of Electrical and Computer Engineering
University of Texas at Dallas, Richardson, TX 75080

***Abstract* – Spatial correlation modeling of wafer-level measurements has been explored in the literature to reduce the test cost and test time. In this project we have implemented wafer level spatial modeling of probe test measurements using gaussian process model with progressive sampling and spatio temporal features. Further one more enhancement is added to handle in-lot variation using wafer-based clustering approach. Finally, we demonstrate the results on the given dataset and also show that the added enhancement improves the prediction error when there is a shift in the process across the wafers in lot.**

I. Introduction

In semiconductor manufacturing transistor sizes are scaled down to produce smaller and high-performance ICs. Recent sub nanometer ICs are prone to process variation and manufacturing defects which might result in devices not meeting its specifications. Hence ICs are tested to ensure their desired behavior within the specification limits specified in the data sheet. The widely acceptable industrial approach to test RF ICs is specification-based testing, which involves verifying sequentially all the performances of the device against the specification limits. Automatic test equipment(ATE) performs the desired tests on the Circuit under test(CUT) to decide its pass/fail status. Measurement of RF parameters is very complex and require expensive test instruments. It increases the cost of ATE and also results in long testing time due to the sequential approach of testing.

Various methodology has been studied in literature to reduce the test cost and research on modelling spatial measurement correlation has shown great promise in capturing wafer-level spatial variation. The underlying idea is to identify spatial correlation between probe test measurements among the die in a wafer by relying only on subset of die measurements and predict performance outcomes at unobserved die locations. One of the proven methods of doing this is a regression technique called gaussian process(GP) modelling [1]. Further enhancements to this model has been made which include inclusion of radial feature, a clustering approach to handle discontinuous effects [3]. Two additional key techniques namely progressive sampling and spatio temporal feature inclusion [2] has further improved the prediction accuracy and computational time.

In this project we first implement the above-mentioned GP model along with the progressive sampling and spatio temporal feature techniques. Progressive sampling aims at improving the statistical information available in the sample based on which a spatial correlation model is constructed. Spatio-temporal feature inclusion, on the other hand, extends the concept of correlation modelling across wafers, asserting that the spatial variation observed on a wafer reveals useful information for other wafers in the same lot. Further we group wafers of similar trend in measurements using a clustering-based approach and build a spatio-temporal correlation model for each of the groups which captures variation of a parameter across the group of wafers in a lot as a function of die coordinates and wafer time-index.

The paper is presented as follows. Section II discusses the models used in our project. Section discusses the implementation flow and V discusses the experimental results, VI concludes.

II. Models used

In this section, we briefly discuss the models we use in the implementation of our project. We first discuss the Gaussian process model, then we go on to explain the two enhancements to the GP model namely progressive sampling and spatio temporal feature.

A. Gaussian Process Model

Gaussian process regression modelling [4] is an inductive regression approach targeted at extrapolating a function over a Gaussian random field based on limited data observations. It is a collection of random variables, any finite number of which exhibits a joint Gaussian distribution. A Gaussian process is fully specified by its mean function and its kernel-based covariance function. Consider a training set D of n_t data points, $D = \{(x_i, f(x_i)) | i = 1, \ldots, n_t\}$ where x denotes an input vector (Here, the input vector is the Cartesian coordinate of a die denoted as x = [x, y]) and f(x) is the output (herein, a measurement value). Accordingly, a Gaussian process can be

viewed as a group of random variables f(x_i) with joint Gaussian distribution:

$$f(x_1),\ldots,f(x_n) \sim N(0,K) \quad (1)$$

where element K_{ij} of the covariance matrix 'K' is the covariance between values f(x_i) and f(x_j).
The default kernel function is given by

$$k(\mathbf{x}_i,\mathbf{x}_j) = \exp\left(-\frac{1}{2l^2}|\mathbf{x}_i-\mathbf{x}_j|^2\right) \quad (2)$$

The basic assumption here is that if vectors x and x′ are similar, then f(x) and f(x′) should be similar, too. The covariance function K(x, x′) returns a measure of the similarity of x and x′ and encodes how similar f(x) and f(x′) should be. Substituting the squared exponential covariance function into the definition of the Gaussian process, we arrive at a Gaussian process formulation as:

$$f(x) \sim GP(0,K(x,x')) \quad (3)$$

For a new data point with input x_*, the predictive distribution of the output $f_*(x_*)$ can be computed by using conditional distributions of the joint Gaussian distribution:

$$f_*|X,\mathbf{t},\mathbf{x}_* \sim \mathcal{N}(\mathbf{k}_*^\top K^{-1}\mathbf{t}, \quad k(\mathbf{x}_*,\mathbf{x}_*) - \mathbf{k}_*^\top K^{-1}\mathbf{k}_* \quad (4)$$

where **X** is the input training matrix, t is a training output vector, x_* is a test point, k_* = K(X, x_*) which is the kernel evaluation between the test point and all training instances, K is the matrix of the kernel function evaluated at all pairs of training points and k(x_*, x_*) is the variance of the kernel function at test point x_*.

To avoid overfitting, a technique known as regularization is often employed in decision-theoretic empirical risk minimization. This is incorporated in the GP model by adding additive noise σ. Equation (3) is updated to include additive noise

$$y = f(\mathbf{x}) + \varepsilon \sim \mathcal{GP}(0, k(\mathbf{x},\mathbf{x}') + \sigma_n^2\delta_{\mathbf{x},\mathbf{x}'}) \quad (5)$$

The predictive distribution results in:

$$f_*|X,\mathbf{t},\mathbf{x}_* \sim \mathcal{N}(\mathbf{k}_*^\top(K+\sigma_n^2 I)^{-1}\mathbf{t}, \quad k(\mathbf{x}_*,\mathbf{x}_*) - \mathbf{k}_*^\top(K+\sigma_n^2 I)^{-1}\mathbf{k}_* \quad (6)$$

resulting in a point prediction for new observations of $f_* = k_*^T (K + \sigma_n^2 I)^{-1} t$. This constrains the fitted model to avoid extreme predictions

B. Progressive Sampling

An iterative progressive sampling approach to select training samples which better represent the spatial variation pattern across a wafer is explained in [2]. The ability of the GP model to provide a confidence level can be leveraged for all predicted samples in each iteration. Algorithm 1 outlines the proposed progressive sampling approach.

1. Randomly select n' samples on the wafer as initial training set: $S = \{\mathbf{x}_1|t_1,\cdots,\mathbf{x}_{n'}|t_{n'}\}$
2. Build spatial GP model using set S and predict values and confidence at unobserved die locations (set U)
3.1 For each $\mathbf{x}_i$ in U, calculate $\mathbf{d}_i = min\{|\mathbf{x}_i - \mathbf{x}_j|^2, \forall \mathbf{x}_j \in S\}$
3.2 Select location $\mathbf{x}_i$ which has highest variance and maximum Euclidean distance from current training set
3.3 Add $\mathbf{x}_i$ to the set S and remove it from set U and obtain corresponding true value $\mathbf{t}_{x_i}$
3.4 Repeat 3.1-3.3 until k locations are added to the training set
4. Augment the training set $S = \{S, \mathbf{x}_{h_1}|t_{h_1},\cdots,\mathbf{x}_{h_k}|t_{h_k}\}$
5. Repeat steps 2-4, until stopping criterion is reached

Algorithm 1: Progressive sampling of information-rich training locations in spatial correlation modeling

C. Spatio Temporal feature

Wafers within the same lot, will exhibit similar intra-wafer spatial variation and will also exhibit time-dependent inter-wafer variation [2]. Essentially the conjecture is, *"A single spatio-temporal model learned from samples across all wafers in a lot, could be more accurate than individual models learned from the same samples for each wafer"*.

An advantage of using Gaussian process regression is the ability to apply a Gaussian process over any arbitrary index set. To accommodate time dependence of wafers in the Gaussian process model, we can simply update the coordinates from x = [x, y] to include a time feature t: x = [x, y, t] Applying Gaussian process regression over this space will result in a model that takes time dependent variation into account.

Combining the two new enhancements to wafer-level correlation modelling is also a plausible direction. Specifically, it is possible that the accuracy of a spatio-temporal model can be improved by employing progressive sampling.

III. IMPLEMENTATION

In this section we illustrate the implementation steps involved in this project:

1) For each of the probe test apply k-means clustering to one wafer and find k clusters. The same clusters are taken for rest of the wafers.
2) For each cluster, take same training points for say 2.5% of devices in the cluster.
3) Append all the training points of the first cluster across all the wafers.

4) Instead of training a single gaussian process model across all the wafers, group the wafers with similar values and train for each group of wafers.

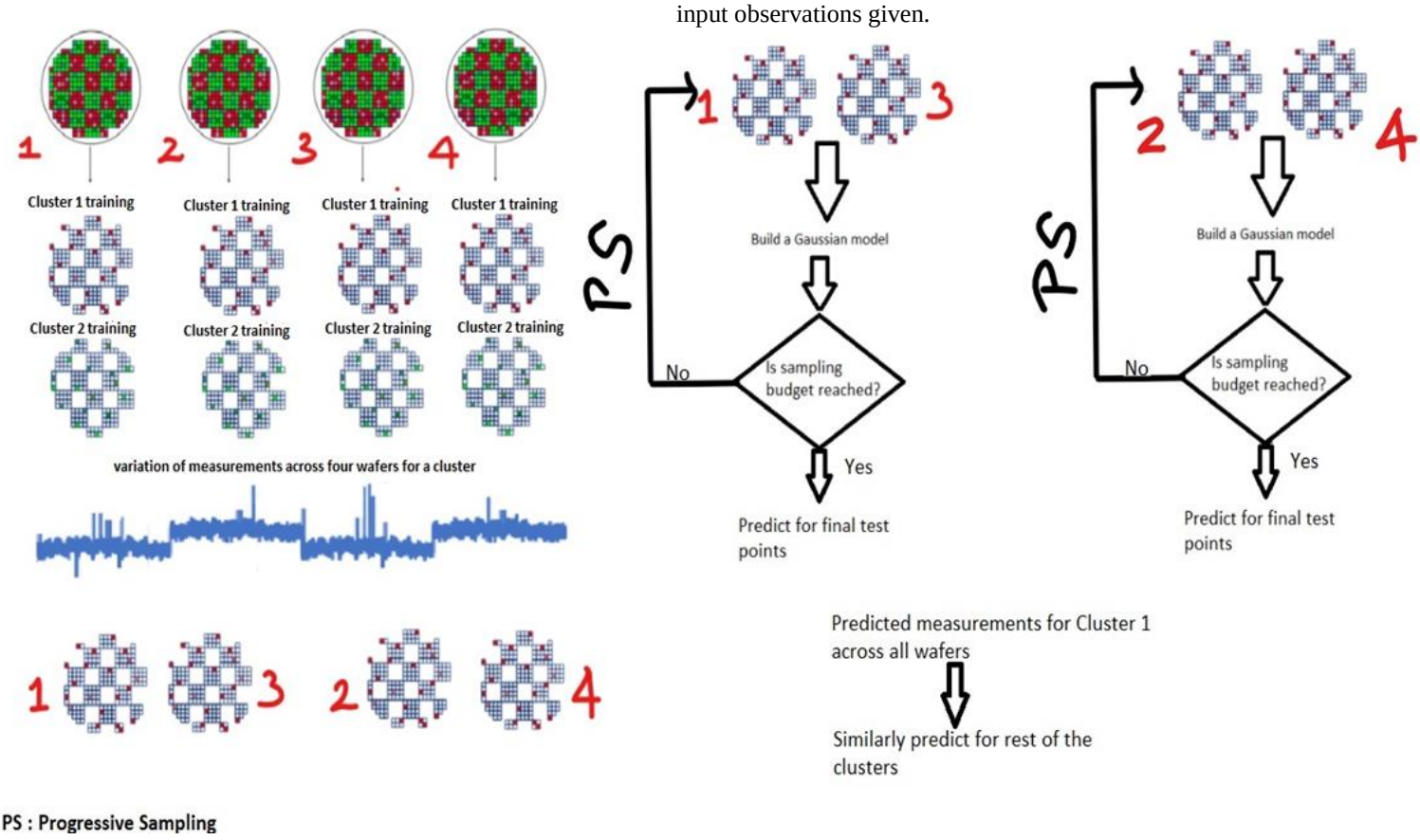

Fig. 1: Flow diagram of the Implementation

5) Model is trained using a gaussian process model with k-fold cross validation for each group of wafers.
6) Predict the values for rest of the locations in the first cluster across all the wafers.
7) Apply progressive sampling by selecting test points with highest variance and adding it to the training set increasing its size by 2.5%
8) Repeat steps 4 to 6 till training set size of 10% of devices in the first cluster is reached
9) Repeat the steps 2 to 8 for all the clusters across the wafers.
10) Repeat steps 1 to 9 for all the probe tests.

As shown in the figure 1, Consider four wafers with discontinuous effects in each of them. Each of the wafers is divided into two clusters using intra-wafer k-mean clustering in the same points across the wafers. As seen in the figure there is process variation across the four wafers. Instead of building a single temporal model across the four wafers, we group wafers having similar trend of measurements using inter-wafer k-mean clustering. Mean and standard deviation is calculated for same initial training points across each of the wafers and given as an input to k-mean model. The k-mean model then clusters the wafers into groups as based on the input observations given.

As shown in the figure wafers 1and 3 and wafers 2 and 4 are grouped together since they have similar trend of measurements values. Now we build a Gaussian model for each of the two groups with progressive sampling approach. Once sampling budget is reached prediction is done for the final test points. Similarly, steps are repeated for rest of the clusters.

IV. EXPERIMENTAL RESULTS

Here we evaluate the results of the methods implemented. The probe test measurements dataset consists of 102 probe tests for a total of 23 wafers and each wafer consists of around 2300 devices. From these experiments we expect to show that if there is a shift in process variations across the wafers within a lot applying inter-wafer based clustering approach with gaussian process model and progressive sampling and spatio temporal feature enhancements should reduce the prediction error of probe test measurements.

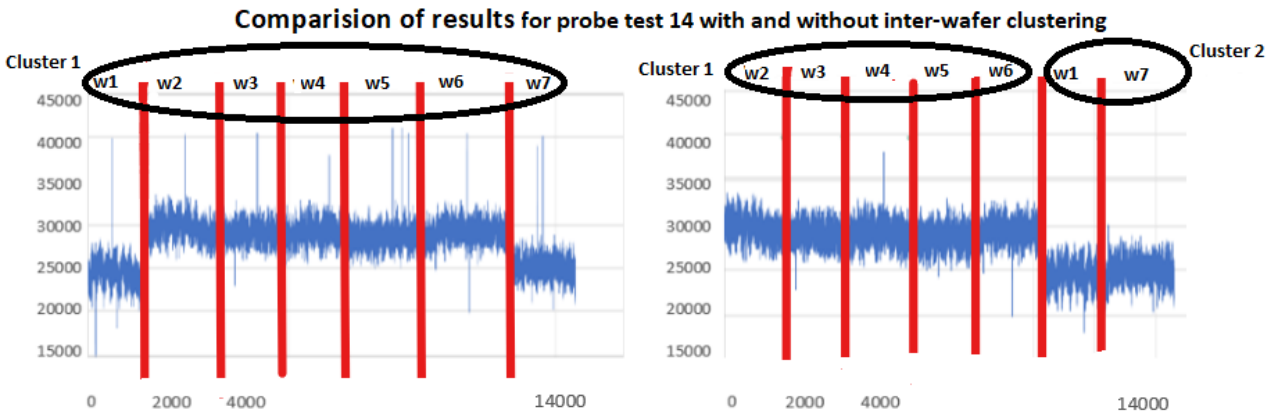

Fig 2a: Actual probe measurement values w/o inter -wafer clustering

Fig 2b: Actual probe measurement values w/ inter -wafer clustering

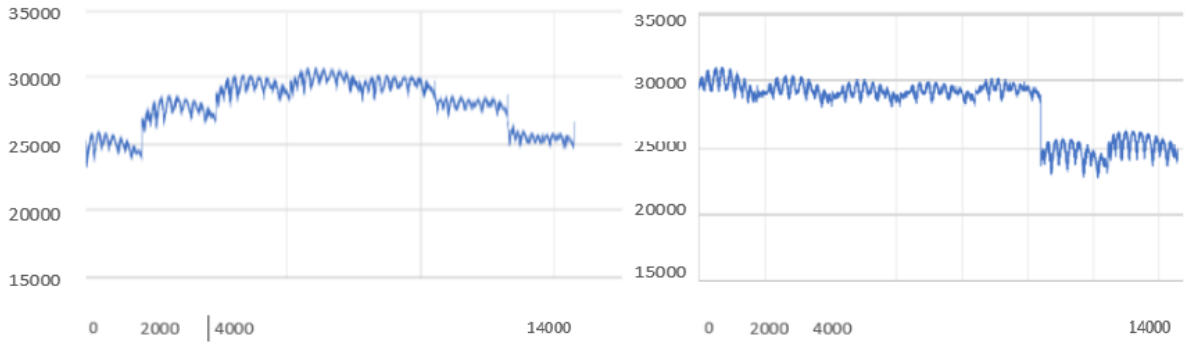

Fig 2c: Predicted probe measurement values w/o wafer clustering inter -wafer clustering

Fig 2d: Predicted probe measurement values w/ inter -wafer clustering

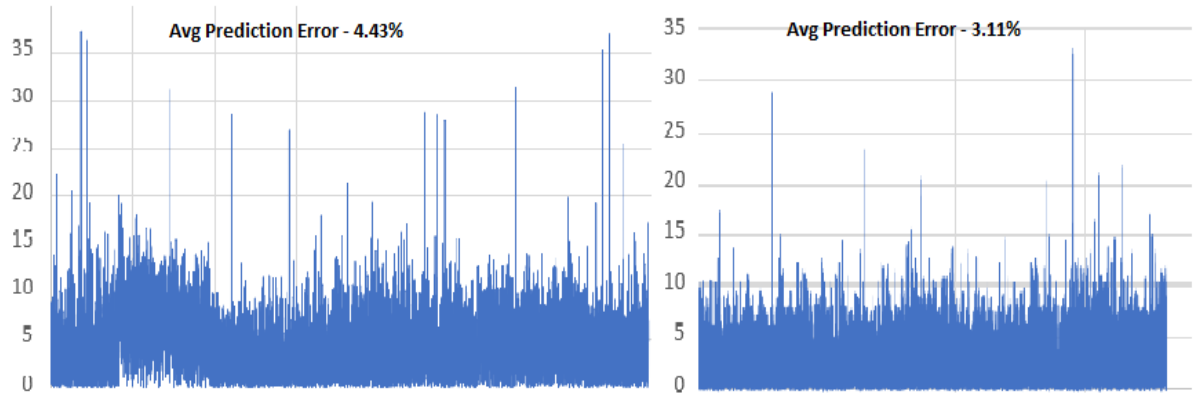

Fig 2e: Average prediction error w/o wafer clustering

Fig 2f: Average prediction error w/ wafer clustering

First, we remove the outliers from each of the probe test measurements. The number of clusters, k, and the shape of the clusters are computed during the pretraining stage, using only the first wafer. Once the k clusters are known, we use a randomly chosen training sample of 2.5% devices from each cluster across all the wafers in order to train the spatial correlation models. Mean and standard deviation is calculated for all initial training points across each of the wafers and using k-means clustering wafers are grouped. In the first increment, the same random locations are chosen on each cluster across groups of the wafers, in order to help the model capture time dependent correlation. Subsequent sample increments are chosen across all the wafers in each group from the first cluster in the lot. Based on the prediction confidence level of the spatio-temporal correlation model constructed using all previous samples from all the wafers in each group for the first cluster. Once the overall sampling budget of 10% is reached (i.e. after 4 iterations), use the spatiotemporal correlation model for each group of wafers to predict this parameter for the remaining 90% of die on all wafers in each group.

In order to assess the accuracy of each model, we compare the values for the predicted die locations (90%) to the actual probe test outcomes and we capture the discrepancy using the absolute relative error metric: $\epsilon = |t' - t|/t$ where t is the probe test outcome for a die, t' is the corresponding predicted value for that die and Specification Range is the range of that measurement across the wafer after outlier removal using a $\pm 3\sigma$ filter.

We compare the predicted probe measurements for probe test 14 without any clustering as implemented in [2] with predicted probe test measurements with inter-wafer clustering. As we see in Fig 2 since there is discontinuous process variation across wafers and since a single gaussian process model is built across all wafers, the predicted probe measurements try to follow the trend for all the wafers which results in less prediction accuracy with average prediction error being 4.43%. But with inter-wafer clustering wafers w2, w3, w4, w5, w6 have similar variation and are grouped together and wafers w1 and w7 are grouped together and we build separate gaussian model for each of the two groups. The resulting prediction values are much more accurate in this case with average prediction accuracy being 3.11%.

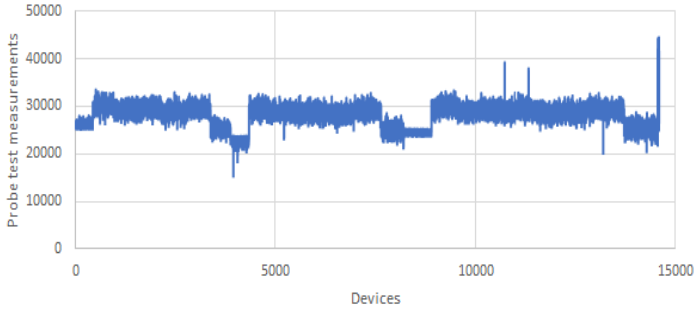

Fig 3a: Actual probe test measurements of probe test 14 w/o inter and intra-wafer clustering

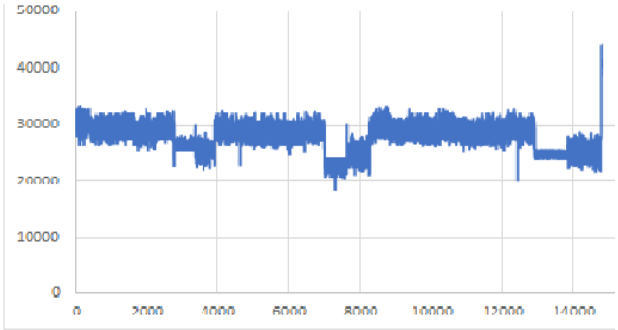

Fig 3b: Actual probe test measurements of probe test 14 w/ inter and intra wafer clustering

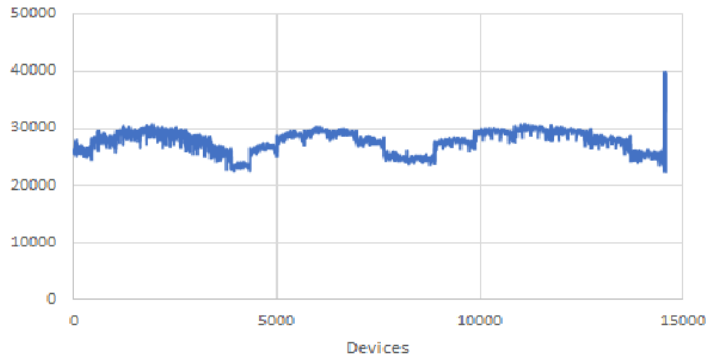

Fig 3c: Predicted probe test measurements of probe test 14 w/o inter and intra-wafer clustering

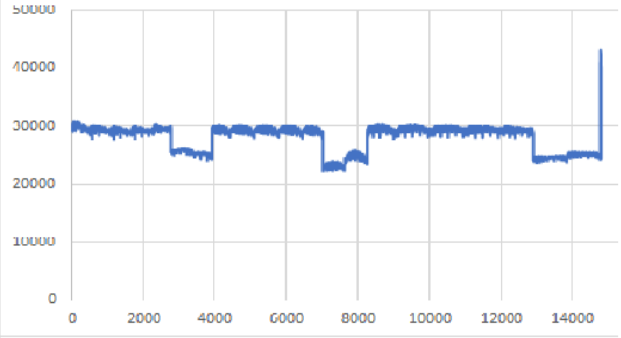

Fig 3d: Predicted probe test measurements of probe test 14 w/ inter and intra-wafer clustering

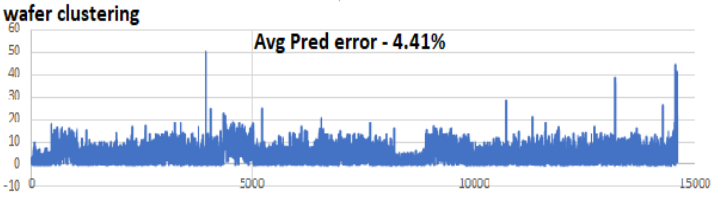

Fig 3e: Prediction error for probe test 14 w/o inter and intra-wafer clustering

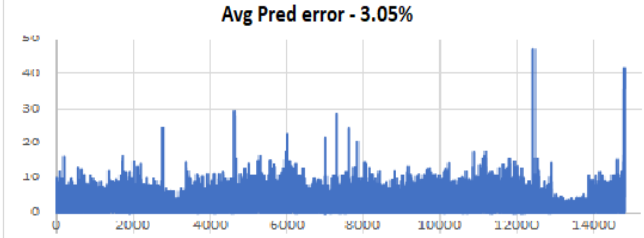

Fig 3f: Prediction error for probe test 14 w/ inter and intra-wafer clustering

Similarly, we compare the predicted probe measurements for probe test 14 with intra-wafer clustering to handle any discontinuous process variation within the wafer, with predicted probe test measurements with inter-wafer and intra-wafer clustering. As we see in Fig 3 the average prediction error with GP model built with only intra wafer clustering is 4.41% and the average prediction error is 3.05% with GP model built with both intra and inter wafer clustering.

One more inference from these results is that since for the probe test 14 there is not much process variation within the wafer, building a GP model with only intra wafer-based clustering doesn't improve the prediction accuracy by much. As observed the GP model along with progressive sampling and spatio temporal feature inclusion and inter and intra-wafer clustering shows improved prediction accuracy and so we run the same method for all the probe tests and plot the results.

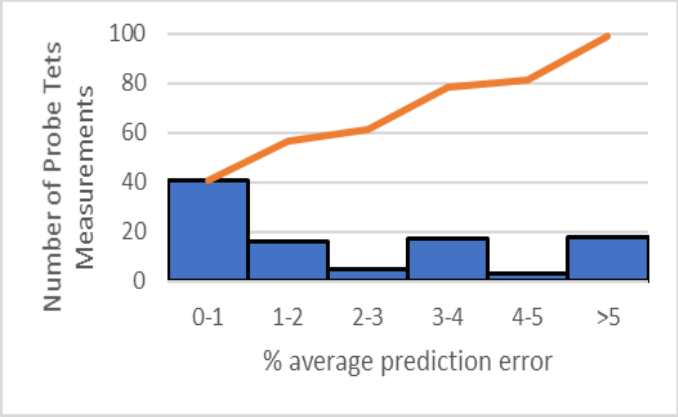

Fig 4: Average prediction error across all the tests

Plotting the average prediction error for all the probe tests with GP model along with progressive sampling and spatio temporal features and with inter-wafer and intra-wafer clustering we can see from the graph that for over 80% of the devices the prediction error is less than 4%. Fig 5 shows the sorted average prediction error for 80% of the probe tests.

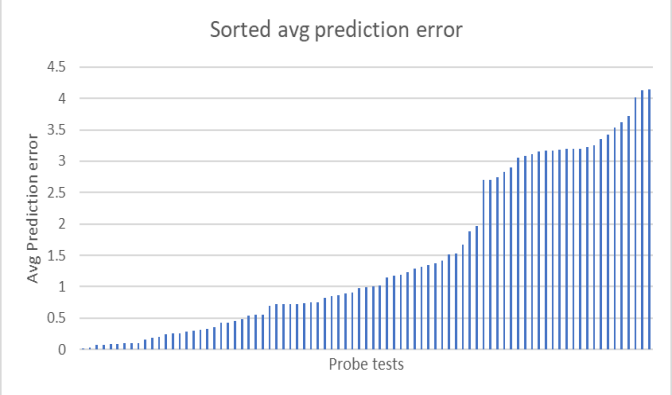

Fig 5: Sorted average prediction error for 80% of probe tests.

We also note that improved model doesn't add any additional complexity during production testing as compared to the model presented in [2] since the decision about choosing the next of set of training points for all the clusters in each group of wafers after every iteration of progressive sampling is made offline and only information ATE needs for each iteration is next set of training points for each cluster in every group of wafers.

V. Conclusion

In this project we first implemented wafer spatial modelling using Gaussian process model along with key enhancements namely progressive sampling and spatio-temporal feature inclusion [2]. Further to handle within wafer and in-lot process variation we added a intra and inter-wafer clustering-based approach to the previous model. The results were compared for one of the probe tests between the two models and improvement in average prediction accuracy was shown. Further we showed that the average prediction error for all the probe tests for the improved model is less than 4% for over 80% of the devices.

References

[1] N. Kupp, K. Huang, J.M. Carulli, and Y. Makris, "Spatial estimation of wafer measurement parameters using Gaussian process models," in IEEE International Test Conference, 2012, pp. 1 – 8.

[2] A. Ahmadi, K. Huang, S. Natarajan, J. M Carulli, and Y. Makris, "Spatio-temporal wafer-level correlation modeling with progressive sampling: A pathway to HVM yield estimation," in IEEE International Test Conference, 2014, pp. 1–10.

[3] K. Huang, N. Kupp, J.M. Carulli, and Y. Makris, "Handling discontinuous effects in modeling spatial correlation of wafer-level analog/RF tests," in Design, Automation & Test in Europe Conference, 2013, pp. 553 – 558.

[4] C.E. Rasmussen and C.K.I. Williams, Gaussian Processes for Machine Learning, MIT Press, 2006.