Proceedings of the 2023 Winter Simulation Conference
C. G. Corlu, S. R. Hunter, H. Lam, B. S. Onggo, J. Shortle, and B. Biller, eds.

MODELING MULTIVARIATE RELATIONS IN MULTIBLOCK SEMICONDUCTOR MANUFACTURING DATA USING PROCESS PLS TO ENHANCE PROCESS UNDERSTANDING

Geert van Kollenburg
Richard Verhoeven
Mike Holenderski
Nirvana Meratnia

Eindhoven University of Technology
Groene Loper 3
Eindhoven, 5612AE, NETHERLANDS

Daniele Pagano

STMicroelectronics s.r.l.
Stradale Primo Sole, 50
Catania, 95121, ITALY

ABSTRACT

The complexity of manufacturing process data has made it more challenging to extract useful insights. Data-analytic solutions have therefore become essential for analyzing and optimizing manufacturing processes. Path modeling, also known as structural equation modeling, is a statistical approach that can provide new insights into complex multivariate relationships between process variables from different stages of the manufacturing process. The incorporation of expert process knowledge and subsequent interpretation of model results can facilitate communication between stakeholders, promoting lean manufacturing and achieving the sustainability goals of Industry 5.0. This paper describes the use of a path modeling algorithm called Process Partial Least Squares (Process PLS) to gain new insights into the relationships between equipment data from several machines within the semiconductor manufacturing process. The methods used in this study can assist manufacturers in understanding the relations between different machines and identify the most influential variables that may be used to develop soft-sensors.

1 INTRODUCTION

The semiconductor industry plays a vital role in modern society, with increasing demand for smaller and more powerful computer chips and electronic devices. Computer chips are made in hundreds at a time on silicon wafers. The manufacturing process involves multiple stages, including wafer fabrication, deposition, lithography, and etching (Timings 2021). Each stage involves multiple process steps and each step may be dependent on the performance of the previous steps. Variations in any of the hundreds of steps can affect the final product performance, yield, and reliability (Melhem et al. 2015). The complexity and dependency of the manufacturing processes, together with the abundance of process data has resulted in an increasing demand for data analytic solutions to analyze and optimize these processes.

Statistical models and machine learning algorithms can provide valuable insights into the manufacturing processes. Yet most data-analytic applications do not model the relationships between the many process steps, but instead focus on predicting specific outcomes like product quality or yield (Biegel et al. 2022; Sanchez-Marquez and Vivas 2020; Dupret et al. 2005) in order to control the manufacturing processes. Path modeling, also known as structural equation modeling (SEM), is a valuable approach that can identify multivariate relationships between process variables from various steps of the manufacturing process (van Kollenburg et al. 2020; Hair Jr et al. 2021).

Path modeling enables the identification of key factors that contribute to the overall variability of the manufacturing processes. By incorporating all relevant variables and dependencies between manufacturing

steps, path models can reveal new insights into the complex relationships between process variables and their impact on the manufacturing process (Vinzi et al. 2010). Path modeling can assist manufacturers in identifying the root causes of process variations, implementing corrective actions, and identifying areas for improvement. Real-time process optimization will become possible when models identify strong relationships between specific process variables and a defect-causing situation later in the process (Arteaga and Ferrer 2002).

Path modeling is particularly useful in process analytics as they allow for the incorporation of expert process knowledge in an intuitive manner. This facilitates communication between data-analysts and manufacturing operators and such explainability is critical if models are to be integrated into personnel's routine work (Meindl et al. 2021; Cagliano et al. 2019). The inclusion of knowledge of process experts and subsequent interpretation of model results provides a unique opportunity to promote lean manufacturing (Tortorella et al. 2019). Collaboration between humans and data-analytic solutions is a core aspect of the sustainability goals specified for Industry 5.0 (Breque et al. 2021). Investing in a working environment, where workers can interact with data-analytics in an intuitive way can lead to increased value creation in manufacturing processes (Cifone et al. 2021; Senoner et al. 2022).

The aim of the work presented in this paper was to learn more about the complex connections between distinct steps in the semiconductor manufacturing process. This paper explains the process of obtaining features from the equipment data, analyzing the data with a path modeling method called Process PLS (van Kollenburg et al. 2021), and evaluating different model configurations to understand the relationships between machines. The research is presented in a way that allows the methods to be applied to other manufacturing processes, which is why the term 'machine' will be used throughout to describe parts of the processes otherwise called 'tools' or 'chambers' and so forth.

The remainder of this paper is organized as follows. The next section discusses related work. In Section 3 the data that was used will be explained. That Section also includes an overview of the Process PLS algorithm and illustrates multiple model specifications. Section 4 presents the results of the analyses using the model specifications discussed. The paper ends with a discussion and future outlook in Section 5.

2 RELATED WORK

Data-analytic applications are common in manufacturing industry (Moldovan et al. 2017). While many applications focus on quality predictions (Köksal et al. 2011 provide an overview), statistical process control has also been standard practice for a long time (Spanos 1992). Path models on the other hand have mostly been used to analyze company-level indicators. Examples include analysis of environmental practice and manufacturing performance (Tseng et al. 2008), sustainable manufacturing practices (Vinodh and Joy 2012), front-end product design (Withanage et al. 2012), corporate governance mechanisms (Fei et al. 2015) and marketing strategies (Sarstedt et al. 2022). To the best of our knowledge, path models have not yet been applied to evaluate the interrelations between various part of the manufacturing process itself.

Path models have recently been used to analyze industrial chemical production processes (van Kollenburg et al. 2020; Offermans et al. 2021), leading to new insights into relations between process variables that led to reduced production costs. Sensors in industrial processes provide time series. In process industry, sensors may measure different bits of material at each point in time as the materials flows through the machines. If properly aligned, each data point of a sensor can be related to all other data points related to the same bit of material (Offermans et al. 2021). As such, the data is well-suited for use in correlational methods like path models.

In semiconductor manufacturing processes, products reside at each machine for some time. This means that many sensors each produce a time series variable per wafer. Multivariate time series can be analyzed with statistical models like ARIMA (Hamilton 2020), upcoming data points can be predicted with deep learning models like LSTM (Hochreiter and Schmidhuber 1997) and anomaly detection can be done with ensemble methods (Trardi et al. 2022). These methods, however, neither consider the multiple-machine

nature of the manufacturing process to model the interrelations between the various sub-processes, nor predict specific features in other time series.

To the best of our knowledge, no path modelling extensions for multi-block time series have yet been developed to accommodate multiple target blocks (Gu and Van Deun 2019). To make use of existing path models, the time series variables must be transformed in such a way that correlations between sets of data become meaningful.

3 MATERIALS AND METHODS

3.1 Data

Historical equipment data from seven machines within a semiconductor manufacturing process was provided by STMicroelectronics s.r.l. The seven machines will be referred to as Machine 1 through 7, in the order that wafers pass through them. The data consisted of readings from 151 sensors which are distributed over the seven machines, providing information on the production of over 2000 wafers, which all followed the same production recipe. While details about the manufacturing steps cannot be disclosed, the goal of the research was to relate specific observations at one machine to observations at other machines. For instance, identifying whether an extreme value in Variable *A* of Machine 1 is related to high variability of Variable *B* on Machine 3.

As a pre-processing step, features of the time series variables were extracted before further analysis. From each variable, the average value (avg), minimum value (min), maximum value (max) and the standard deviation (std) were extracted. This means that each of the 151 time series was transformed into 4 features, with each feature having a single value per wafer. This transformation is illustrated in Tables 1 and 2. The data was labeled in the format `M_V_feature`, where `M` indicates the machine at which the measurement was done ranging from `M1` to `M7`, `V` indicates the sensor number, ranging from `V1` to `V151` and `feature` is the label for either the `avg`, `min`, `max`, or `std`.

Table 1: Illustration of the original time series that was used to create the data set used in further analyses. Labels are given in format Machine_Variable. Each of the 151 sensors produced a variable and each wafer thus had observations on these 151 variables. Data is shown for the first and last wafer.

Wafer 1			.	Wafer 2186		
M1_V1	.	M7_V151	.	M1_V1	.	M7_V151
.774	.	.499	.	.746	.	.166
.519	.	.753	.	.64	.	.818
.283	.	.041	.	.069	.	.31
.786	.	.19	.	.038	.	.767
.294	.	.196	.	.092	.	.05
.	.	.	.	.	.	.
.	.	.	.	.	.	.
.	.	.	.	.	.	.
.315	.	.973	.	.821	.	.538
.619	.	.239	.	.981	.	.193
.137	.	.583	.	.859	.	.376
.324	.		.	.327	.	
.149	.		.	.733		
.522	.		.			

The columns in Table 1 represent sensor readings observed for each wafer. Each row in that Table indicates a time point during which the wafer was in the respective Machine. Also illustrated is the fact that not all time series were of equal length. In Table 2 each column represents a feature from the time series and each row represents the observed value of that feature for a given wafer. The boxed number is

the minimum value observed in Variable 151 for Wafer 1. The encircled number indicates the maximum value observed in Variable 1 for Wafer 2186. Each feature (min, max, avg, std) was named according to the machine and the variable they originated from in the format 'Machine_variable_feature'. For example, the feature `M1_V1_max` lists for each wafer the maximum value observed in the time series `M1_V1` that was collected at Machine 1. In the remainder of this paper, the term 'features' will refer to the data used in the statistical analyses, as represented in Table 2.

Table 2: Features extracted from the time series that were used in statistical analyses.

	M1_V1_max	M1_V1_min	.	M7_V151_min	M7_V151_avg}	M7_V151_std
Wafer 1	.786	.137	.	.041	.434	.322
.	.	.	.	.	.	.
.	.	.	.	.	.	.
.	.	.	.	.	.	.
Wafer 2186	.981	.038	.	.05	.402	.282

We excluded features that had more than 15% missing values or that had (near-)zero variance. Near-zero variance may be related to both rare events and to uninformative data, but for the analyses presented below, such features cannot be used. Then, we removed 26 wafers that had missing data for any of the remaining features. To prepare the data for analysis, we standardized all features to have a mean of zero and a variance of one. The resulting data set consisted of 562 features describing the 151 time series during the production of 2186 wafers. This data set was used for subsequent analyses. The analyzed data and code to reproduce the results presented below may be provided to the interested reader upon reasonable request.

3.2 Process PLS

Partial least squares (PLS), also known as Projection to Latent Structures (Wold 1982), is a family of multivariate regression techniques used to evaluate the relationships between two or more blocks of data. PLS models can handle large number of correlated predictor variables. A high correlation between predictors is called collinearity and is problematic for most traditional regression techniques because it leads to rank-deficiencies in covariance matrices.

Standard PLS regression only models the relation between two blocks of data. This could be used to predict the data of, say, Machine 4 from the data of Machine 3 (See Figure 1a). Merging multiple blocks into one larger block makes it possible to predict, for example, data of Machine 5 from all the data observed in the machines preceding it (i.e. Machines 1, 2, 3, and 4 as shown in Figure 1b), but this approach does not model the contributions of individual machines. There are also multi-block PLS extensions that can handle multiple predictor blocks (Figure 1c), where the relation between each pair of blocks is evaluated separately. The reader is referred to other literature for a discussion on the standard approach for the multiple-predictor models (Biancolillo and Næs 2019).

To study the relationships between the machines, information from one machine should be utilized to predict data from another machine, while itself being predicted by other machines. In essence, any block of data can simultaneously act as both a predictor and a target, with several targets existing within the model. We employ a versatile path model algorithm called Process PLS (van Kollenburg et al. 2021), which provided information on the relations between the data from the various machines in the manufacturing process.

Like other SEM models, a Process PLS model consists of an outer and inner model. The outer model partitions the data into blocks. In the current context, this means that features are grouped according to their respective machines. In Figure 2, the outer model specification is represented by the blue components of the model and the machines as red rectangles. The inner model, represented as the green arrows in Figure 2, is used to specify which connection between machines to include. For the remainder of this paper, all following figures will only present the inner model.

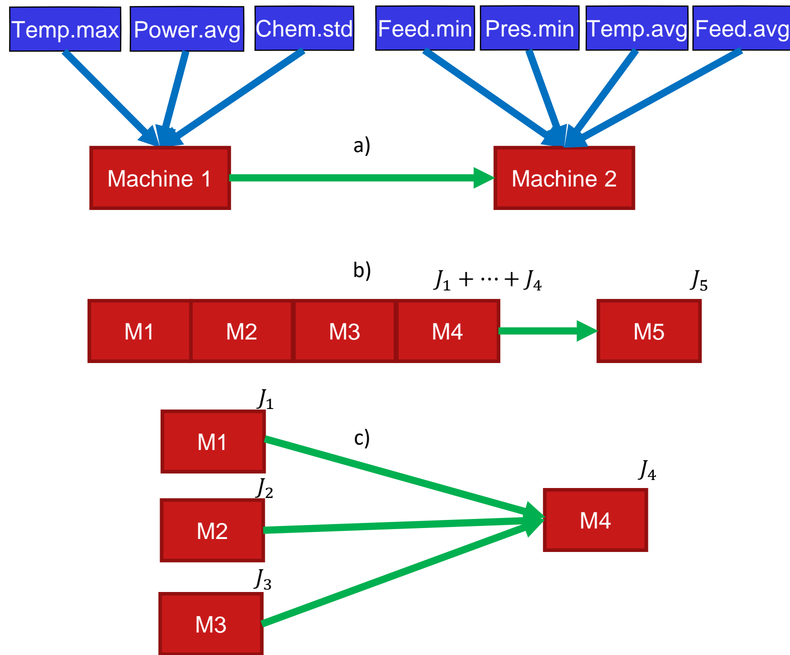

Figure 1: Examples of PLS model specifications with a single target block. The model a) uses data from one machine to predict the data from another machine. Sub-figure b) illustrates forming a single predictor block. Sub-figure c) shows a multi-block specification. J_m indicates the number of columns in the data of block m. Standard PLS regression can only be used to analyse models a) and b).

The Process PLS algorithm has two main steps to find optimal predictions of all target blocks. First the outer model is estimated by constructing lower-dimensional latent variables from the features in each block, based on their relationships with features of target blocks. The proportion of variance in the features of block m that is contained in the lower-dimensional representation of that block is indicated by R^2_m. In single-target models, these R^2 values are equivalent to the R^2 values obtained in standard PLS regression (see Figure 1a). While often called 'explained variance', R^2 in PLS is better interpreted as the proportion of variance extracted from the data to ensure optimal prediction. In the current application, the data consists of manually extracted features. It can be expected that several features will not be informative. If much data is redundant, interpreting absolute R^2 values, being a proportion of all variance in a block, should be avoided. To ensure interpretability, Process PLS has a second step.

In the second step of Process PLS, one PLS regression model is estimated for each target block. All blocks that predict a particular target block are used as predictors in the model. The primary result of this step is the explained variance P^2 (Rho-squared). For a given block m, the explained variance is calculated as the sum of all (partial) explained variances of every predictor block n, where n represents the predictors of m. In other words,

$$P^2_m = \sum^n P^2_{m.n}$$

represents how much of the total variance in the lower-dimensional representation of the target block m can be predicted by (latent variables of) other blocks. Conceptually, $P^2_{m.n}$, as a partial explained variance, can be compared to the square of a regression coefficient. For instance, a $P^2_{m.n}$ value of .5 implies a regression effect of block n on m equivalent to $\sqrt{.5} = .71$. While this analogy is strictly conceptual in case of multiple latent variables per block, it may offer readers anintuitive grasp of the meaning of the model results. Please refer to the foundational paper of Process PLS for details (van Kollenburg et al. 2021).

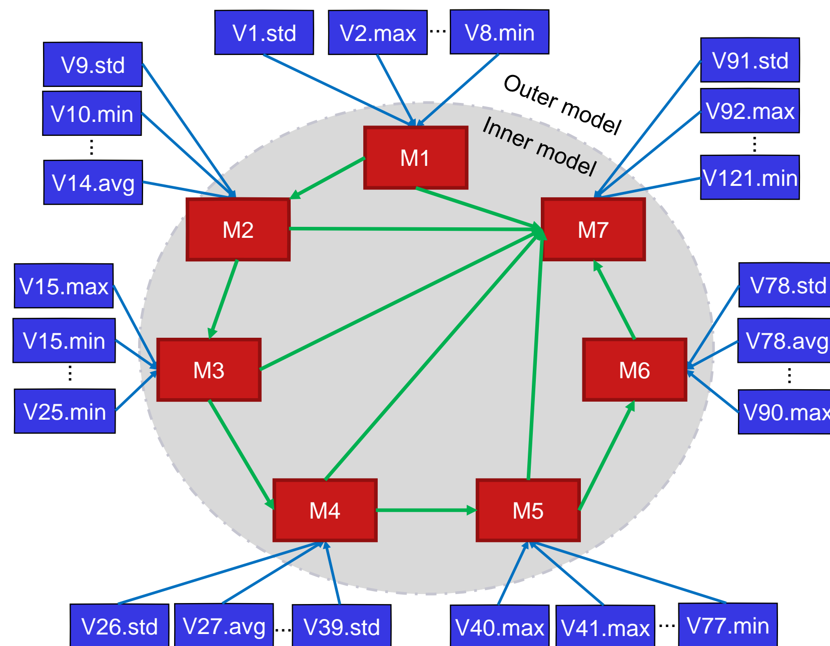

Figure 2: Example of a Process PLS model. Data from each machine is used to predict the following machine and each machine is used to predict data from the last machine. The outer model is represented in blue, the machine representations are in red, and the green arrows represent the inner model.

Explaining a small yet interesting portion of the data could be more valuable than explaining a large portion of redundant data. As a result, P^2 values serve as more robust indicators of the strength of relationships between the blocks compared to R^2 values. Suppose the R^2 value of block m is $R^2_m = .2$. Next to limited relation with other data, factors such as collinearity or the presence of uninformative features might have caused the R^2 value to be low. By calculating the P^2_m and $P^2_{m.n}$ values, which are a proportion of the R^2_m value, we can obtain a more robust measure to determine if there is predictable data in block m.

3.3 Model Specification

The inner model of a Process PLS model can be specified in multiple ways. One strategy is to first model the flow of the process and then, with a different model, include relationships between blocks that one wants to explore. Modelling the flow of a wafer through the manufacturing process can be done with the inner model specification as shown in Figure 3a. Here, the model assumes that data from each machine is only dependent on the machine before it and that relations between machines can be explained by this flow. From a substantive point of view, it may be unlikely that such a Markovian process is true, yet it remains valuable for illustration purposes.

The Process PLS specification just discussed is not identical to having multiple two-block models. The optimization procedure of Process PLS optimizes block representations to be as predictive as possible. For example, Machine 2 (M2) functions as a target in the relation between M1 and M2. In a two-block model with only M1 and M2, M2 would be optimally predicted. But since M2 functions as a predictor of M3, the lower-dimensional representation of M2 is optimized to be as predictive as possible for M3. We stress that if the goal of ones research is to find optimal predictions of each block of data, path modelling is not the most optimal option. Multiple dedicated prediction models will be better suited for that purpose.

Next to modelling the flow of the wafers, it is also of interest to model the relationship between machines that are further apart in the process. For the current application, the main interest was in predicting the data from the last machine from each of the machines preceding it. The model specification related to this

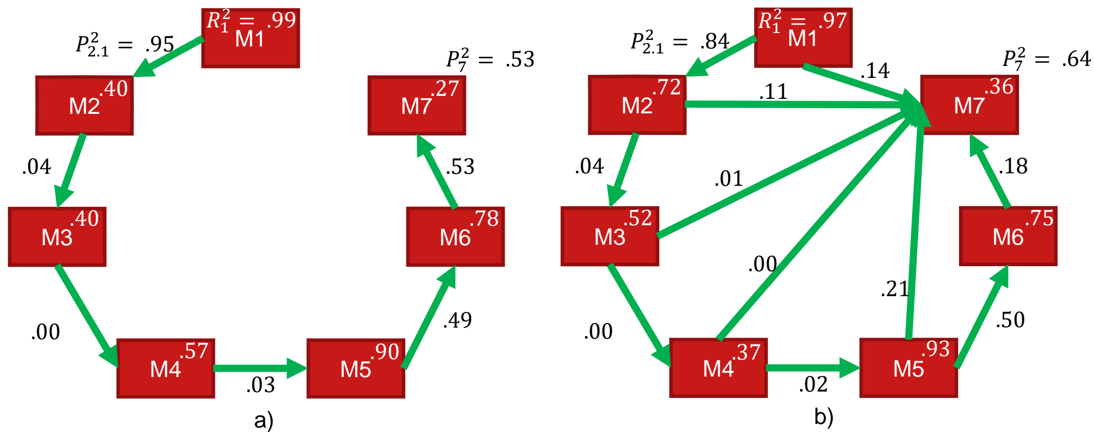

Figure 3: The left inner model specification represents the flow of a wafer through the manufacturing process. The right specification includes relations of all machines with the final machine. R^2_m values are shown within each rectangle. The $P^2_{m.n}$-values shown near the arrows indicate how much of the extracted information in the target block can be predicted from the predictor block.

is shown in Figure 3b. The multiple-predictor, multiple-target structure is illustrative of the added value of path modelling. That is, the many relations are estimated simultaneously and detailed information can be gathered about which parts of the process are most strongly related to other parts, conditional on the other relations. We also note that this method of model specification is especially interesting if the final block of data contains quality and/or yield measurements (which it does not in the current case). In this way, the relations between multiple machines and product quality can be studied simultaneously, which can benefit process control optimization.

Various other model topologies beyond those presented in this paper are possible, each of which may offer unique insights into the process. We instead present a data-driven approach to identify a simplified model that highlights only the strongest relationships. This can be done by specifying an inner model that estimates all effects between machines (shown in Figure 4), and removing the inner model connection with the lowest P^2 value. This results in a model with one fewer connection. We repeating this process iteratively until all remaining P^2 values exceeded .10. This value was subjectively chosen and is comparable to a regression coefficient of .33. If in any iteration, a machine had no outgoing or incoming arrows, we removed the data of that block from the data set.

4 RESULTS

The first model (Figure 31) indicated a strong relation between data from Machine 1 and Machine 2. Of the 40% variance extracted from the data related to Machine 2 (i.e., $R^2_2 = .40$), 95% could be predicted from features from Machine 1 (i.e., $P^2_{2.1} = .95$). From a substantive point of view, this supports prior expectations expressed by process experts and was a first confirmation that the algorithm performed as intended. The model also showed a strong relation between (features from) Machines 5 and 6 ($P^2_{6.5} = .49$) and between Machines 6 and 7 ($P^2_{7.6} = .53$). There was virtually no relation between the features from the other Machines. Since the absolute values of R^2 are not meaningful in the current context (see last paragraph of Section 3.2), these values are only shown in the figures, but not will not be discussed. The discussions will only focus on the regression effects P^2.

In the second model (Figure 3b), most effects representing the flow were highly similar to those found in the first model. The notable exception is that the effect of Machine 6 on Machine 7, $P^2_{7.6}$, dropped from

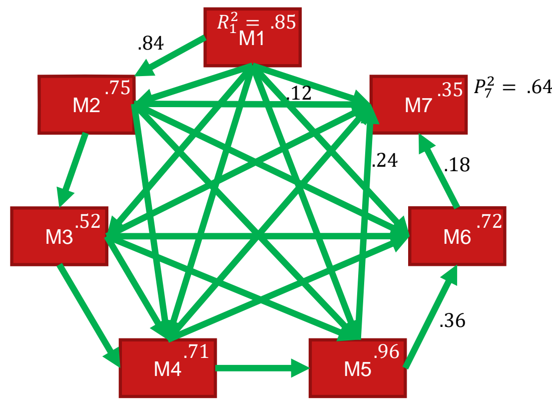

Figure 4: Inner model with each machines predicting all following machines. This model was used to iteratively find a simplified model in which all P^2 values are at least .10. All inner model effects that exceeded .10 in this model specification are shown in the graph. All effects are provided in Table 3.

.53 in the first model to .18 in the second model. This decrease in $P^2_{7.6}$ means that part of the data that could be explained from Machine 6 can also be explained by data from other machines. The total proportion of variance of Machine 7 that could be explained in the second model calculated as $P_7^2 = \sum_{n=1}^{6} P^2_{7.n} = .64$ (i.e., the sum of all effects on Machine 7) is higher than in the first model, where P_7^2 was .53. In the second model, more data is being used to predict Machine 7, leading not only to a higher explained variance, but also showing details on the contributions of each machine to the predictions of Machine 7.

The data-driven approach to find a simplified model started with the fully connected model presented in Figure 4. The inner model effects are provided in Table 3. Due to rounding, some effects are presented in the table as being equal (e.g., multiple entries being .00). The procedure of estimating the model and removing the connection with smallest effect and then re-estimating the new model took 16 iterations before all effects, $P^2_{m.n}$, exceeded .10. The final simplified model is shown in Figure 5.

Table 3: Inner model coefficients ($P^2_{m.n}$) from the fully connected model.

	Predictors (n)					
Target(m)	M1	M2	M3	M4	M5	M6
M2	.84					
M3	.03	.04				
M4	.03	.03	.00			
M5	.04	.04	.00	.01		
M6	.08	.08	.00	.00	.36	
M7	.09	.12	.01	.01	.24	.18

The effects in the simplified model provide interesting details on the interrelations between the machines. The strong relations between Machines 1 and 2 is still represented well in this model, as is the relation between Machines 5 and 6. Model simplification hardly affected the the ability to predict the data of Machine 7 ($P_7^2 = .63$). The effect of Machines 1, 2, 4 and 6 on Machine 7 were quite similar, indicating that each machine has a distinct relation with different parts of the data in Machine 7. This is crucial for studying possible solutions for process control.

Detailed investigation of the results showed that `M1_V7_max`, `M2_V76_avg`, `M5_V125_min` `M6_V139_avg` and `M7_V151_min` where the most important features in their respective machines in terms of their contribution to the variance of the lower-dimensional representations. Details are provided in Table 4, with some description of what these features represented. To give an example, `M2_V76_avg` is a feature (the average) extracted from the time series Variable 76. Variable 76 reports the readings of a

temperature sensor during the time a wafer is in Machine 2. The average value of this sensor contributed 21.1 times more to the Process PLS model than the other features from Machine 2 contributed on average.

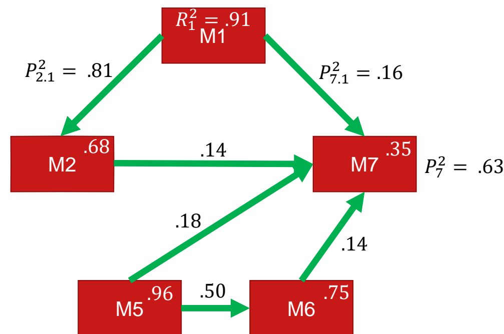

Figure 5: Simplified inner model found through iteratively removing inner model connections from a fully specified model (Figure 4 until all P^2 values were at least .10. As no relations with Machines 2 and 3 reached the threshold, the respective data blocks were removed from the model.

Table 4: Anonimized names of the most important features of each block in terms of contribution to the (simplified) Process PLS model. Feature contributions are an indication of how much variance a feature contributes to the variance of the lower-dimensional representations. The relative contribution is calculated as the contribution of this feature divided by the average contribution of all features in its block.

Feature Name	**Relative Contribution**	**Description**
M1_V7_max	8.2	The maximum value of a voltage
M2_V76_avg	21.1	Average temperature
M5_V125_min	7.4	Minimum of a certain flow
M6_V139_avg	6.1	Average pressure
M7_V151_min	20.6	Minimum of a certain speed

It is possible to evaluate which features are most important for predicting certain blocks of data. For example, the single strongest predictor for the data in Machine 2 was feature M1_V36_std, which indicates the variability in temperature on Machine 1 during manufacturing. The minimum of a certain flow, M5_V125_min, was the single most important predictor of the data observed in Machine 6. For Machine 7 the variability of a different flow, represented by feature M5_V126_std was the strongest predictor. Without providing too many details, it may be clear that this type of information is extremely valuable for predictive process control and obtaining valuable insights for root-cause analysis.

5 DISCUSSION

This paper showcases how path modeling can serve as a potent instrument to study relationships between process data across production lines. This approach helped us reveal correlations between feature sets, leading to a deeper understanding of the intricate interactions in the production process. Path models are interpretable, and their simple topology and the ease with which models can be specified, creates an intuitive representation of the process. By simplifying the models, we were able to identify the essential relationships.

The findings obtained from our analyses can be applied towards creating soft-sensors or virtual metrology. It is imperative to identify the crucial connections between features in order to determine which input variables should be used. Subsequently, these features can be utilized to forecast the quality of end-products

(van Kollenburg et al. 2022), leading to production optimization, cost reduction, and improvement in overall product quality.

Our research showed the importance of addressing co-linearity in process data in complex systems like production lines. We illustrated the methodology using a basic feature selection method, which likely missed key relations between the time series. Directly integrating time series into a path modeling framework would better capture complex relations and result in more accurate models. Until this becomes possible, a priori filtering is needed.

6 CONCLUSION

The multitude of possible relations between time series features makes multivariate cross-correlation infeasible for our purposes. Path modelling allowed us to use domain-specific insights to enhance the effectiveness of our models and better understand the relations in the data. Although our current model has shown some success in addressing our research questions, there is a need for a better-fitting model that can account for complex interactions between variables not anticipated by domain experts. A promising direction for future research is the extension of path modeling to tensor-based regression models (Liu et al. 2021), which may overcome the limitations of manual feature selection by capturing higher-order relationships in the data.

For future research, we suggest developing a neural-network-based path modeling combined with tensor data representations from manufacturing processes. Such methods could improve model performance by learning complex relationships between different types of data. Incorporating expert knowledge into the model may become even more crucial to ensure interpretability. Neural network models, capable of adapting to changes and discovering hidden data patterns, could offer more robust solutions to manufacturing challenges.

ACKNOWLEDGMENTS

This work was in part supported by ECSEL Joint Undertaking, under grant agreement No 826589. The authors express their gratitude to Paola Giuffre, Caterina Genua and Daniele Li Rosi for their consultation with respect to the manufacturing process and data presented in this paper. OpenAI's ChatGPT (GPT-4) was used to proofread the presented text, checking for spelling and grammar. In no way were algorithms used to create original content or to generate ideas.

REFERENCES

Arteaga, F., and A. Ferrer. 2002. "Dealing With Missing Data in Mspc: Several Methods, Different Interpretations, Some Examples". *Journal of Chemometrics: A Journal of the Chemometrics Society* 16(8-10):408–418.

Biancolillo, A., and T. Næs. 2019. "The Sequential and Orthogonalized Pls Regression for Multiblock Regression: Theory, Examples, and Extensions". In *Data handling in Science and Technology*, Volume 31, 157–177. Elsevier.

Biegel, T., N. Jourdan, C. Hernandez, A. Cviko, and J. Metternich. 2022. "Deep Learning for Multivariate Statistical In-Process Control in Discrete Manufacturing: A Case Study in a Sheet Metal Forming Process". *Procedia CIRP* 107:422–427.

Breque, M., L. De Nul, and A. Petridis. 2021. "Industry 5.0: Towards a Sustainable, Human-Centric and Resilient European Industry". *Luxembourg, LU: European Commission, Directorate-General for Research and Innovation*.

Cagliano, R., F. Canterino, A. Longoni, and E. Bartezzaghi. 2019. "The Interplay between Smart Manufacturing Technologies and Work Organization: The Role of Technological Complexity". *International Journal of Operations & Production Management* 39(678):913–934.

Cifone, F. D., K. Hoberg, M. Holweg, and A. P. Staudacher. 2021. "'Lean 4.0': How Can Digital Technologies Support Lean Practices?". *International Journal of Production Economics* 241:108258.

Dupret, Y., E. Perrin, J. Grolier, and R. Kielbasa. 2005. "Comparison of Three Different Methods to Model the Semiconductor Manufacturing Yield". In *IEEE/SEMI Conference and Workshop on Advanced Semiconductor Manufacturing 2005.*, 118–123. IEEE.

Fei, G. C., A. M. Rasli, and S. S. Xuan. 2015. "Modeling the Heterogeneity of Corporate Governance Mechanisms Across Industries: A Multi-Group Analysis Using PLS Path Modeling". *Journal of Contemporary Issues and Thought* 5:50–65.

Gu, Z., and K. Van Deun. 2019. “Regularizedsca: Regularized Simultaneous Component Analysis of Multiblock Data in R”. *Behavior Research Methods* 51:2268–2289.

Hair Jr, J. F., G. T. M. Hult, C. M. Ringle, and M. Sarstedt. 2021. *A Primer on Partial Least Squares Structural Equation Modeling (Pls-sem)*. Sage Publications.

Hamilton, J. D. 2020. *Time Series Analysis*. Princeton, New Jersey: Princeton University Press.

Hochreiter, S., and J. Schmidhuber. 1997. “Long Short-term Memory”. *Neural Computation* 9(8):1735–1780.

Köksal, G., I. Batmaz, and M. C. Testik. 2011. “A Review of Data Mining Applications for Quality Improvement in Manufacturing Industry”. *Expert Systems with Applications* 38(10):13448–13467.

Liu, J., C. Zhu, Z. Long, Y. Liu et al. 2021. “Tensor Regression”. *Foundations and Trends® in Machine Learning* 14(4):379–565.

Meindl, B., N. F. Ayala, J. Mendonça, and A. G. Frank. 2021. “The Four Smarts of Industry 4.0: Evolution of Ten Years of Research and Future Perspectives”. *Technological Forecasting and Social Change* 168:120784.

Melhem, M., B. Ananou, M. Djeziri, M. Ouladsine, and J. Pinaton. 2015. “Prediction of the Wafer Quality with Respect to the Production Equipments Data”. *IFAC-PapersOnLine* 48(21):78–84.

Moldovan, D., T. Cioara, I. Anghel, and I. Salomie. 2017. “Machine Learning for Sensor-Based Manufacturing Processes”. In *2017 13th IEEE International Conference on Intelligent Computer Communication and Processing (ICCP)*, 147–154. IEEE.

Offermans, T., L. Hendriks, G. H. van Kollenburg, E. Szymańska, L. M. Buydens, and J. J. Jansen. 2021. “Improved Understanding of Industrial Process Relationships through Conditional Path Modelling with Process Pls”. *Frontiers in Analytical Science* 1:721657.

Offermans, T., E. Szymańska, G. H. van Kolllenburg, L. M. Buydens, and J. J. Jansen. 2021. “Automatically Optimizing Dynamic Synchronization of Individual Industrial Process Variables for Statistical Modelling”. *Computers & Chemical Engineering* 152:107402.

Sanchez-Marquez, R., and J. M. J. Vivas. 2020. “Multivariate Spc Methods for Controlling Manufacturing Processes Using Predictive Models–a Case Study in the Automotive Sector”. *Computers in Industry* 123:103307.

Sarstedt, M., J. F. Hair, M. Pick, B. D. Liengaard, L. Radomir, and C. M. Ringle. 2022. “Progress in Partial Least Squares Structural Equation Modeling Use in Marketing Research in the Last Decade”. *Psychology & Marketing* 39(5):1035–1064.

Senoner, J., T. Netland, and S. Feuerriegel. 2022. “Using Explainable Artificial Intelligence to Improve Process Quality: Evidence from Semiconductor Manufacturing”. *Management Science* 68(8):5704–5723.

Spanos, C. 1992. “Statistical Process Control in Semiconductor Manufacturing”. *Proceedings of the IEEE* 80(6):819–830.

Timings, J. 2021. “6 Crucial Steps in Semiconductor Manufacturing”. *https://www.asml.com/en/news/stories/2021/semiconductor-manufacturing-process-steps, accessed 24 March 2023*.

Tortorella, G. L., R. Giglio, and D. H. Van Dun. 2019. “Industry 4.0 Adoption as a Moderator of the Impact of Lean Production Practices on Operational Performance Improvement”. *International Journal of Operations & Production Management* 39(678):860–886.

Trardi, Y., B. Ananou, P. Tchatchoua, and M. Ouladsine. 2022. “Ensemble Machine Learning Algorithms for Anomaly Detection in Multivariate Time-Series”. In *2022 International Conference on Control, Automation and Diagnosis (ICCAD)*, 1–6. IEEE.

Tseng, M.-L., W.-W. Wu, Y.-H. Lin, and C.-H. Liao. 2008. “An Exploration of Relationships between Environmental Practice and Manufacturing Performance Using the Pls Path Modeling”. *WSEAS Transactions on Environment and Development* 4(6):487–502.

van Kollenburg, G., R. Bouman, T. Offermans, J. Gerretzen, L. Buydens, H.-J. van Manen, and J. Jansen. 2021. “Process Pls: Incorporating Substantive Knowledge into the Predictive Modelling of Multiblock, Multistep, Multidimensional and Multicollinear Process Data”. *Computers & Chemical Engineering* 154:107466.

van Kollenburg, G., M. Holenderski, and N. Meratnia. 2022. “Value Proposition of Predictive Discarding in Semiconductor Manufacturing”. *Production Planning & Control*:1–10.

van Kollenburg, G. H., J. van Es, J. Gerretzen, H. Lanters, R. Bouman, W. Koelewijn, A. N. Davies, L. M. Buydens, H.-J. van Manen, and J. J. Jansen. 2020. “Understanding Chemical Production Processes by Using Pls Path Model Parameters as Soft Sensors”. *Computers & Chemical Engineering* 139:106841.

Vinodh, S., and D. Joy. 2012. “Structural Equation Modeling of Sustainable Manufacturing Practices”. *Clean Technologies and Environmental Policy* 14(1):79–84.

Vinzi, V. E., L. Trinchera, and S. Amato. 2010. “PLS Path Modeling: From Foundations to Recent Developments and Open Issues for Model Assessment and Improvement”. In *Handbook of Partial Least Squares: Concepts, Methods and Applications*, edited by V. Esposito Vinzi, W. W. Chin, J. Henseler, and H. Wang, 47–82. Berlin, Heidelberg: Springer Berlin Heidelberg.

Withanage, C., T. Park, T. T. H. Duc, and H.-J. Choi. 2012. “Dynamic Partial Least Square Path Modeling for the Front-end Product Design and Development”. *Journal of Mechanical Design* 134(10):100907.

Wold, H. O. 1982. “Soft modelling: the Basic Design and some Extensions”. In *Systems under Indirect Observation, Part II*, 36–37. North Holland.

AUTHOR BIOGRAPHIES

GEERT VAN KOLLENBURG is Assistant Professor at Eindhoven University of Technology in the Netherlands. He has worked with national and international market leaders to develop data-driven solutions for sustainable chemical and semiconductor industries. His expertise spans statistical process control, structural equation modelling, chemometrics, and more. His dedication to sustainability and expertise in statistics and machine learning have led to innovative approaches for predictive and prescriptive analytics. His email address is g.h.v.kollenburg@tue.nl and website is https://research.tue.nl/nl/persons/geert-van-kollenburg.

RICHARD VERHOEVEN is a University Researcher and IT Developer at the Department of Mathematics and Computer Science at the Eindhoven University of Technology in the Netherlands. He has worked together with national and international partners on the topics of component based systems, wireless sensor networks, internet of things and hardware-in-the-loop simulations. His email address is p.h.f.m.verhoeven@tue.nl and website is https://research.tue.nl/nl/persons/richard-verhoeven.

DANIELE PAGANO is Funding Project Manager at STMicroelectronics s.r.l. He has covered various position and responsibilities in Catania Wafer Fab Operations (Lithography, Dry Etching, APC & SPC, Epitaxy, Process Control), Past experiences in collaborative projects like IMPROVE (2012), INTEGRATE (2015), MADEin4 (2022) and nowadays HiCONNECTS, SATURN, IPCEI. He is author and co-author of several publications on journals and international conferences. His e-mail address is daniele.pagano@st.com.

MIKE HOLENDERSKI is an assistant professor at the Department of Computer Science and Mathematics at the Eindhoven University of Technology in the Netherlands. He did his PhD in Computer Science on the topic of multi-resource management in embedded real-time systems. His current research focuses on reliable and trustworthy machine learning for process optimization and control and hybrid data/knowledge driven modelling of high-dimensional data. His e-mail address is m.holenderski@tue.nl and homepage is https://www.tue.nl/en/research/researchers/mike-holenderski.

NIRVANA MERATNIA is Full Professor of Pervasive Computing at Eindhoven University of Technology in the Netherlands. Her research covers various aspects of computing including distributed machine learning and AI, embedded/edge AI, data-driven networking and smart sensor systems. She has been involved in several national and international projects addressing (distributed) computation and intelligence in the context of Internet of Things and Cyber Physical Systems creating societal and economic impacts. Her e-mail address is n.meratnia@tue.nl.