Preprints of the 21st IFAC World Congress (Virtual)
Berlin, Germany, July 12-17, 2020

# Data Science Challenges in Chemical Manufacturing

**Birgit Braun[1], Ivan Castillo[1], Mark Joswiak[1], You Peng[1], Ricardo Rendall[1], Alix Schmidt[1], Zhenyu Wang[1], Leo Chiang[1]*, Brenda Colegrove[2]**

*[1]Chemometrics & AI, Dow, Lake Jackson, TX 77541, USA*
*[2]Plastics Process Characterization R&D, Dow, Lake Jackson, TX 77541, USA*
**Corresponding author, e-mail: HChiang@Dow.com, Tel: +1-979-238 5377*

**Abstract:** Industrial processes are ripe with data and offer countless opportunities for applied data science, machine learning and artificial intelligence. While process automation and control are providing more guidance in normal operating states, the need for data analytics is abundant when dealing with deviations from defined states, aiming at consistent transitions, or exploring new operating states to optimize production. This paper provides a brief overview of some examples, and introduces a real-life case study available to educators to challenge engineering students in preparation for roles in the chemical industry.

*Keywords:* Data-driven modeling, artificial intelligence, education, process modeling

## 1. INTRODUCTION

The chemical industry, amongst most other manufacturing segments, is undergoing a significant transformation guided by concepts referred to collectively as Industry 4.0. The overarching drive of this fourth industrial revolution is towards state-of-the-art automation and seamless data exchange for faster, better informed decision making. Data analytics is an essential component of Industry 4.0 as the means to derive contextualized insight from various disparate data sources. With the increasing capabilities of sensor technologies and robotics, industrial processes are ripe with data and offer countless opportunities for applied data science, machine learning and artificial intelligence to unlock value. While process automation and control are providing more guidance in normal operating states, the need for data analytics is abundant when dealing with deviations from defined states, aiming at consistent transitions, or exploring new operating states to optimize production (Chiang, Russell, and Braatz, 2001 and Chiang, Lu and Castillo, 2017). As systems become more complex, the need for engineers to efficiently extract signal from data increases tremendously, requiring data literacy and analytics knowledge in the next generation of chemical engineering graduates. Given the essential nature of contextualization of analytics results for value generation in chemical processes (Qin and Chiang, 2019), subject matter experts need to be highly engaged or partake in the analysis.

In this paper, we aim to provide selected illustrations of successful Industry 4.0 scenarios with a special focus on data analytics. We will also present a real-life industrial case study that is available to interested parties in the educational field. The purpose of providing this anonymized dataset is to expose chemical engineering students to a problem, which is representative of many challenges that can be solved using advanced data analytics approaches in chemical manufacturing.

## 2. INDUSTRIAL EXAMPLES FOR DATA-DRIVEN MODELING, MACHINE AND DEEP LEARNING

### *2.1 Model Maintenance and Management*

Before the application of data science methodologies to chemical processes, countless successes were documented in the field of analytical chemistry through the application of chemometrics focused on the spectral interpretation and concentration predictions of compounds from spectroscopy and gas chromatography. Dow has been harvesting the value of chemometrics since its early stages, with property and additive predictions from spectroscopy analysis for Dow's polyethylene business being one key application segment. At any given time, more than 50 process analyzers are running continuously to collect spectra, which are used to predict key mechanical properties and additive concentrations for hundreds of product grades to ensure product quality according to specifications irrespective of the producing plant (Petzekatis et al. 2017). Due to the chemical structures of the additives, which include antioxidants, slip agents, fillers, UV-absorbers and similar, the spectrum exhibits significant overlap in key absorbance regions. Partial least square (PLS) modelling to correlate sub-regions of the spectrum with actual concentrations determined using a primary method is an established method to handle these challenges. The resulting models are expected to be valid across various instruments provided they adhere to defined settings. However, model performance monitoring in such a setting is extremely challenging albeit vital for quality assurance and acceptance. Historically, manual collection of grab samples and comparison of precisely aligned spectral prediction with primary analysis was the sole approach for performance assessment. Increased connectivity and computing power now allows for utilization of a much larger data volume to compare defined average spectral prediction with locally collected primary analysis on an aligned grab sample. This

200-fold increase in data volume (Fig.1) provides a much more complete assessment of model prediction weaknesses as illustrated in Fig. 2. Here, the time aligned PLS prediction is shown versus the actual property highlighting a prediction bias for the same product between two production trains (left graph). The solid horizontal lines correspond to the critical delta based on a 2-sided Z-test (population variances are known). This finding now enables an active investigation into possible root causes and solutions to address higher prediction errors specific to a single production train. The right graph in Figure 2 provides a different view of the predicted versus actual property for various products produced on the same train; while the predictions are within the critical delta for statistical difference, some products fall entirely outside the tolerance band. Understanding these performance weaknesses of the model enables the plant to operate with high confidence for products where the model performs well, and rely on more frequent grab samples for products with high prediction error while the model is being updated.

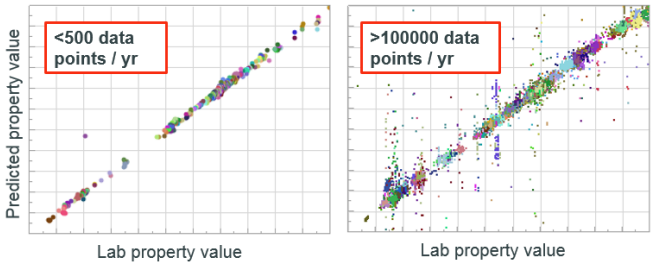


Fig. 1. Predicted versus accurate property value illustrating the data volume in the traditional performance assessment (left) and plant data collection (right). Color coding refers to different product grades.

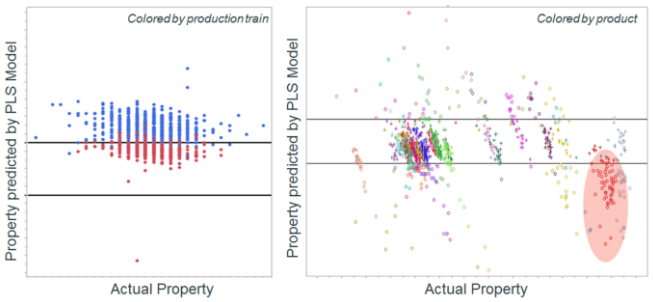


Fig. 2. Predicted versus accurate property for the same product grade produced in two locations (left) and for multiple products produced on the same train (right).

## 2.2 Pellet Shape Classification

Pellet shape is an important quality metric in polyethylene production with impact beyond visual aesthetics. Shapes that deviate strongly from the traditional spherical pellet can cause feeding problems at converters, and shape defects like wispy plastic tails can break off, rapidly oxidize and cause black inclusion in products like sheets or films. Being able to reliably detect pellet shape defects, some of which are shown in Fig. 3, is the basis for adjusting process and pelletizing conditions to maximize quality. Equipment is available on the market to collect silhouette images of pellet samples and extract multiple morphometric factors, but this information is not actionable without reliable classification into shape groups. The determination of the correct metrics and appropriate limits for classification is challenging in a univariate manner. Principal component analysis (PCA) is an unsupervised dimensionality reduction methodology suited in identifying key variables that describe the variability in multivariate datasets. Fig. 4 shows a score plot of a training dataset comprised of six or more morphometric factors for six different pellet shape classes. There is clear clustering present that can be exploited for selecting robust classification criteria. Compared to the original method developed using a univariate selection approach based on subject matter expertise, the PCA enabled classification algorithm reduces misclassification rates significantly, especially in the most quality-critical category, namely tails (Fig. 5). In the training datasets, the misclassification rates were less than 3% for all classes.

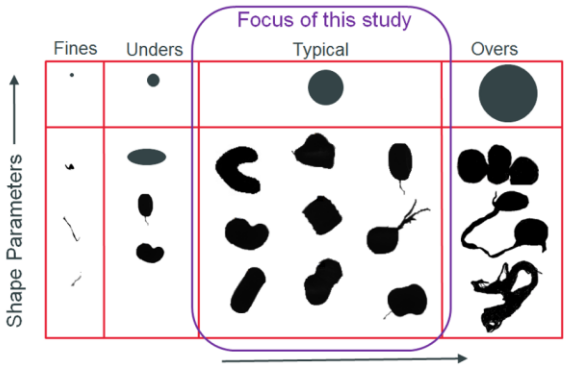


Fig. 3. Typically observed pellets and examples of defects.

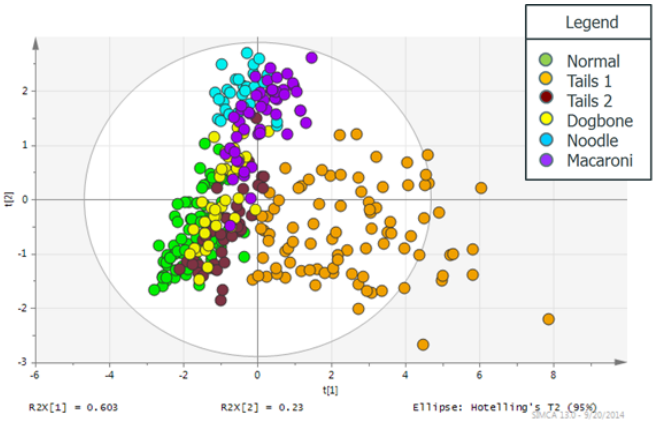


Fig. 4. PCA score plot colored by visually determined pellet shape class.

The classification can be further improved using machine learning and deep learning methodologies as shown in Table 1. Random forest models are well suited for classification problems, especially with categorical variables. Despite the simplicity of the algorithm, it performs very well for both the training and test datasets using the features provided by the analytical instrument. In absence of features, images analysis typically relies on deep neural networks.

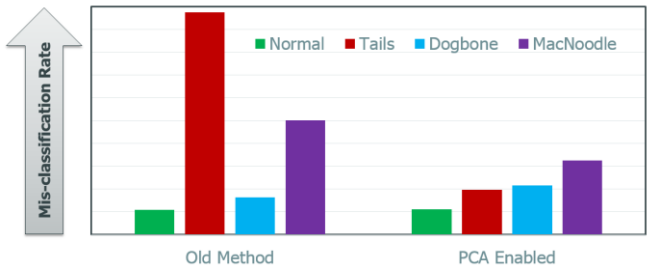


Fig. 5. Misclassification rate for pellet shapes using the incumbent and PCA enabled method.

Multiple architectures were explored and found to provide accurate predictions, which however did not reach the performance of a standard architecture convolutional neural network (VGG-16). The power of transfer learning is highlighted when the VGG-16 architecture is used with pre-trained weights using the ImageNet dataset, and only the weights in the last few layers being adjusted during the training process. The predictions are more robust, which manifests itself in the highest validation and test accuracies (Table 1). For the purpose of this work, accurate predictions are the key focus since Type I and II errors both result in undesired misclassifications into other categories. This work is documented in more detail in Rendall et al. 2018.

Table 1. Classification accuracy for various methods. (PSSD method refers to the PCA enabled method using morphometric features, SDNN is an in-house designed deep neural network architecture).

| Set | PSSD[1] | Random Forests[1] | SDNN | SDNN[2] | Transfer Learning (VGG-16) [2] |
|---|---|---|---|---|---|
| Training | 0.816 | 1 | 0.98 | 0.964 | 0.971 |
| Validation | 0.817 | 0.941 | 0.913 | 0.956 | 0.966 |
| Test | 0.805 | 0.937 | 0.917 | 0.957 | 0.967 |

[1] Approaches based on features
[2] Uses sample augmentation techniques

## 3. INDUSTRIAL CASE STUDY FOR EDUCATORS

The masked data originates from one of Dow's processes. In the selected plant section, which is shown in Fig. 6, impurity accumulation resulted in accelerated catalyst aging (see Fig. 7). The impurity to be predicted is measured at a column overhead in a separation section of the plant, which is the source of 40 process variables (Table 2). The training dataset spans from December 2015 until January 2017, the validation data covers February until October 2017. The dataset has missing values and contains outliers that need to be identified and removed prior to modelling. For the problem to be solved, the process stability and presence of different operating states needs to be analyzed and visualized. A complete list of challenges are included in the instructions provided alongside the dataset, but the main focus is around the development of a reliable and robust prediction model for the impurity concentration. Interpretability is essential to enable root-cause identification and process control opportunities. The intention of making this dataset available to educators is to enable students to practice various data analytics approaches on a real-world dataset. To build a successful model with this dataset, students will need to properly apply preprocessing and visualizion techniques along with variable selection methods. The dataset and detailed instructions can be obtained from Dr. Leo Chiang (HChiang@dow.com).

Fig. 8 shows one of the results for the validation dataset obtained from Dow internal modelling techniques. The prediction accuracy R2 value for this model is 0.698, and the predictions capture most systematic concentration swings. For Dow's purposes, this model proved sufficient as a starting point to interrogate key variables and drive improvement projects. Over the time, the model was revised and further improved.

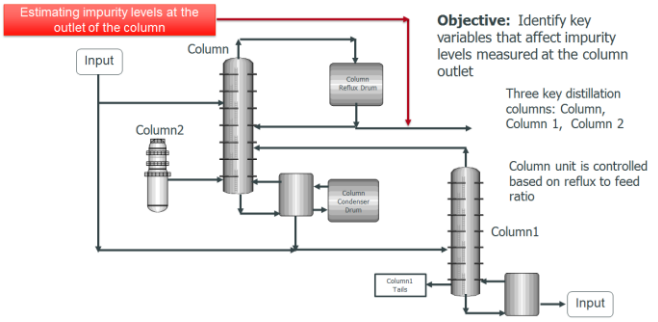


Fig. 6. Block diagram of the plant section considered for the industrial case study for educators.

Table 2. Summary of process variables included in the industrial case study for educators.

| Column Variables | Column1 Variables | Column2 Variables |
|---|---|---|
| x1:Column Reflux Flow | x22: Column1 Base Concentration | x36: Column2 Recycle Flow |
| x2:Column Tails Flow | x23: Flow from Input to Column1 | x37: Column2 Tails Flow to Column |
| x3:Input to Column Bed 3 Flow | x24: Column1 Tails Flow | x38: Column2 Calculated DP |
| x4:Input to Column Bed 2 Flow | x25: Column1 Tray DP | x39: Column2 Steam Flow |
| x5:Column Feed Flow from Column2 | x26: Column1 Head Pressure | x40: Column2 Tails Flow |
| x6:Column Make Flow | x27: Column1 Base Pressure | |
| x7:Column Base Level | x28: Column1 Base Temperature | |
| x8:Column Reflux Drum Pressure | x29: Column1 Tray 3 Temperature | |
| x9:Column Condenser Reflux Drum Level | x30: Column1 Bed 1 Temperature | |
| x10:Column Bed1 DP | x31: Column1 Bed 2 Temperature | |
| x11:Column Bed2 DP | x32: Column1 Tray 2 Temperature | |
| x12:Column Bed3 DP | x33: Column1 Tray 1 Temperature | |
| x13:Column Bed4 DP | x34: Column1 Tails Temperature | |
| x14:Column Base Pressure | x35: Column1 Tails Concentration | |
| x15:Column Head Pressure | | |
| x16:Column Tails Temperature | | |
| x17:Column Tails Temperature 1 | | |
| x18:Column Bed 4 Temperature | | |
| x19:Column Bed 3 Temperature | | |
| x20:Column Bed 2 Temperature | | |
| x21:Column Bed 1 Temperature | | |
| Avg_Reactor_Outlet_Impurity | | |
| Avg_Delta_composition column | | |
| y:Impurity | | |
| Column reflux/feed | | |
| Column make/reflux | | |

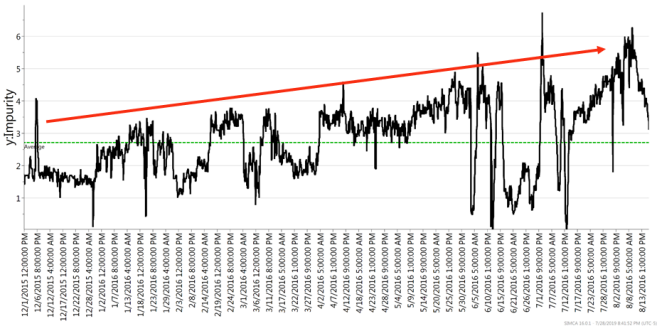


Fig. 7. Impurity concentration over time in training period.

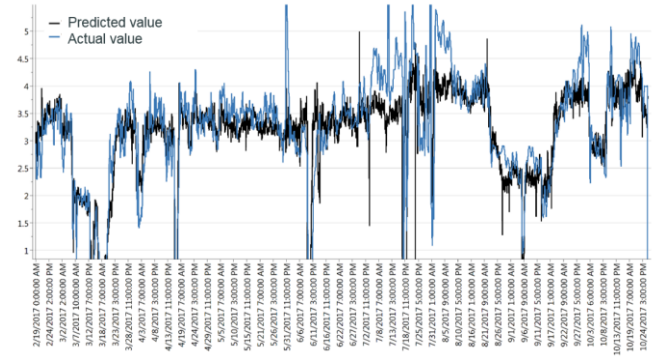


Fig. 8. Predicted (black) and actual (blue) impurity concentrations during validation period.

## 4. CONCLUSIONS

This paper aims at highlighting the broad opportunity space for data analytics in chemical manufacturing, in particular when combined with other Industry 4.0 related efforts. Process and quality data is more readily available and analytics are enabled by more compute power and easier to implement algorithms. We highlight two examples to illustrated how analytics generates value at Dow. Further we introduce an anonymized dataset that is available to educators to serve as illustrative example for students for the types of data analytics challenges that are typically encountered in chemical processes. For more information contact Dr. Leo Chiang (HChiang@dow.com).

## REFERENCES

Chiang, L., Russell, E., Braatz, R. (2001) *Fault Detection and Diagnosis in Industrial Systems*, Springer-Verlag, London.

Chiang, L., Lu, B., Castillo, I. (2017) *Big data analytics in chemical engineering*, Annual Review of Chemical and Biomolecular Engineering, 8:4.1-4.23.

Qin, J. and Chiang, L. (2019) *Advances and opportunities in machine learning for process data analytics*, Computers and Chemical Engineering, 126:465-473.

Petzetakis, N., Braun, B., Hunt, J., Frederick, L., Colegrove, B., Stephenson, S. (2017) *Performance Monitoring of Global Chemometric Models in Manufacturing Plants: New Approaches for Efficient Model Maintenance*, 2017 AIChE Spring Meeting (ISBN: 978-0-8169-1098-4)

Rendall, R., Broadway, M., Lu, B., Castillo, I., Chiang, L., Colegrove, B., Reis, M. (2018) *Image-based Manufacturing Analytics: Improving the Accuracy of an Industrial Pellet Classification System using Deep Neural Networks*, Chemometrics and Intelligent Laboratory Systems, 180: 26-35.