ONLINE ISSN 2069-7430
ISSN-L 1841-4737
PRO LIGNO
www.proligno.ro
Vol. 15 N° 4 2019
pp. 13-23

DATA SCIENCE REVOLUTION FOR SUSTAINABLE BIOMATERIALS ACADEMIC PROGRAMS – A FOCUS ON DATA QUALITY

Timothy YOUNG
University of Tennessee, Center for Renewable Carbon
2506 Jacob Drive, Knoxville, TN 37996-4542
Tel: +01 865 9461119, Fax: +01 865 9461109, E-mail: tmyoung1@utk.edu

Abstract:
The goal of this research effort is to develop a data depository for processing 'real-time automated data fusion' and 'advanced algorithms for data quality improvement,' both essential elements for successful data mining using big data for 'Cloud Manufacturing.' Pragmatic data mining and predictive analytics are the newest initiatives for improved business competitiveness. Predictive analytics can improve manufacturing processes and lower costs of manufactured product. As a critical part of this research, a 'data fusion/quality improvement depository' was developed for the sustainable biomaterials allowing companies to submit anonymous databases for data fusion and data quality improvement. The research advanced Industrie 4.0 by developing advanced algorithms with software interfaces for these industries.

Key words: *data quality; data fusion; data science; Industrie 4.0.*

INTRODUCTION

Predictive analytics, data mining, and the use of *'big data*' are paramount to success in international business and research communities. Data mining and big data are fundamental to '*Cloud Manufacturing*' in the fourth industrial revolution known as '*Industrie 4.0.*', *i.e., "… where computers and automation come together in a new way, with remote connectivity to computer systems equipped with machine learning algorithms that are predictive (*Zhong *et al.* 2017*)*." The sustainable biomaterials industries and related agricultural industries are defined as companies that manufacture: biofuels, bioenergy, lignocellulose products, Nano-biomaterials, wood composites (*e.g.,* particleboard, medium density fiberboard, etc.), engineered wood (*e.g.,* oriented strand board, laminated veneer lumber, cross laminated lumber, etc.), paper and paper products, processing of agricultural products (*e.g.,* rice processing, soybean processing, wheat processing, corn processing, etc.) from renewable feedstock sources. The industries exist in highly competitive markets that are commodity-based, where competitive advantage is sought by lowering the final costs of manufactured product. These industries are important to the U.S. economy, but are facing growing competition from imported products and substitution towards nonrenewable petroleum-derived products, *e.g., PVC flooring, plastic moldings, petroleum fuels, petroleum carbon fibers, etc.* These companies can benefit from '*Cloud Manufacturing*' and the realization of '*Industry 4.0*' where predictive analytics and real-time models are essential to improved decision-making for lowering costs.

Successful data mining is not attainable without digital data integration and data of high quality, *i.e., data of high value*. '*Big data*' is defined relative to its context, *i.e., the standard definition is defined as multiple terabytes or petabytes.* Gandomi and Haider (2015) suggest "*big data is a subjective label attached to situations in which human and technical infrastructures are unable to keep pace with a company's data needs."* In the context of sustainable biomaterials and related agricultural industries, '*big data*' from industrial processes may only be hundreds of gigabytes or less than one terabyte. However, big data of any size, if it has poor quality, it has little value for business improvement from applications of data mining and predictive analytics.

The importance of big data, data mining, and AI is emerging quickly in the sustainable biomaterials industries with the advent of '*Industry 4.0.*' This is exemplified by the Composite Panel Association (CPA) annual meeting in March 2018 for its membership with a focus on data mining and AI for manufacturing.[*] LIGNA 2017, the largest international gathering of sustainable biomaterials industries focused on '*Industry 4.0.*'[†] An important example in the use of big data for these industries is the application of real-time data mining and predictive analytics that are accurate in predicting failures which will directly help manufacturers reduce scrap product and rework (*e.g., reduce remanufacture of off-grade or failing production runs*). An estimate from personal conversations with two wood composite manufacturers is that approximately $1.5 million in losses per mill occur annually due to scrap and rework from typical manufacturing facilities with

[*] https://www.compositepanel.org/news-events/annual-meetings/2018-spring-meeting.html
[†] http://www.ligna.de/en/register-plan/for-journalists/press-infos/press-releases/deutsche-messe-press-releases/ligna-2017-delivers-big-on-innovations.xhtml

annual capacities of approximately 500 mm ft^2.[‡] These losses do not account for additional energy losses and labor costs due to poor quality. Predictive analytics may also help manufacturers diagnose unknown sources of variation from the process (Deming 1986, 1993). Process variation in weight, thickness, solvent applications, drying, etc., create significant costs for manufacturers in that variation influences operational targets. The more variation in a process, the higher the required operational targets for the key inputs required to maintain specification of final manufactured product, which represent additional costs not accounted for in scrap and rework alone (Taguchi 2005).

There is a plethora of literature on data mining and big data (Blanc *et al.* 2008, Gelman *et al.* 2011, Lee *et al.* 2011, Manyika *et al.* 2011, Barton and Court 2012, Kim *et al.* 2012, Rich 2012, Dumbill 2013, Fosso *et al.* 2015, Zhong *et al.* 2016, Deepak and Rani 2018, and many others). However, a review of the literature suggests some gaps in data science research as related to the issue of '*automated data fusion*' which incorporates automated algorithms for '*data quality assessment and improvement*.' Liao *et al.* (2017) did an extensive review of the literature related to '*Industry 4.0*' and found the majority of published research was insufficient in addressing digital integration of data and lack of affordable data mining software. Moreover, the analyses by Liao *et al.* (2017) suggest a huge gap exists between '*Industry 4.0*' laboratory experiments (95.1%) and industrial applications (4.9%). The proposed research effort will reduce this gap by developing algorithms for data fusion and data quality improvement in a '*data depository*' for the sustainable biomaterials and agricultural-based industries. Theorin *et al.* (2015) called data the "*hidden asset in manufacturing,*" while Panetto and Molina (2008) highlighted the lack of utilization of data in manufacturing. Theorin *et al.* (2016) perceived that for future manufacturing systems to be competitive, "*…they need to make better use of plant data, ideally utilizing all data from the entire plant. Low-level data should be refined to real-time information for decision making, to facilitate competitiveness through informed and timely decisions*." Theorin *et al.* (2015) estimated that 85% of all manufacturing data are unstructured and not useful for rapid decision making in high throughput production facilities.

Most sustainable biomaterials and related agricultural processing companies gather real-time digital data from process sensors across programmable logic controller (PLC) networks that are stored in real-time data warehouses. Such data are retrieved by operational personnel to monitor and assess the stability of the process. A parallel information stream is also typically maintained by destructive testing or quality control (QC) laboratories where critical product quality data are gathered periodically throughout the production cycle *e.g., tensile strength, modulus, water absorption, protein content, starch content, etc.* Due to the time required for destructive testing or QC assessments, the time gap in data retrieval from the laboratories to operations personnel may be several hours. In the sustainable biomaterials industries, large quantities of product are produced on high throughput presses during this time gap. Predicting real-time quality attributes between the time gaps from periodic laboratory samples may be invaluable to operations personnel in avoiding the manufacture of defective product, or off-grade product. Data mining may also help diagnose sources of unknown variation in the process. Cost savings can be significant if corrective action on sources of variation occurs which leads to variation reduction of key inputs such as weight. For example, a reduction of 0.5% in operating weight targets for a 500 mm ft^2 capacity wood composite mill may result in cost savings of almost $2 million dollars annually.[§] Lowering operating weight is challenging since this reduces board density which in turn lowers strength (Fig. 1). But through data mining and modeling, other variables in the process can be identified that positively impact strength resulting in an ability to lower density while maintaining strength. Lower density saves materials costs because less feedstocks are needed to manufacture the composite. Less feedstock usage helps for more efficient utilization of natural resources, while lighter weight saves on transportation costs.

[‡] Personal telephone conversations were held with two plant managers in August 2018 that managed a medium density fiberboard and particleboard mill in the southern United States, respectively. The plant managers wish to remain anonymous, but can be contacted by reviewers or representatives from USDA to verify the personal communications.
[§] Personal telephone conversations with a plant manager of a particleboard mill in the southeast U.S. in August 2018. The plant manager wishes to remain anonymous, but can be contacted by reviewers or representatives from USDA to verify this personal communication.

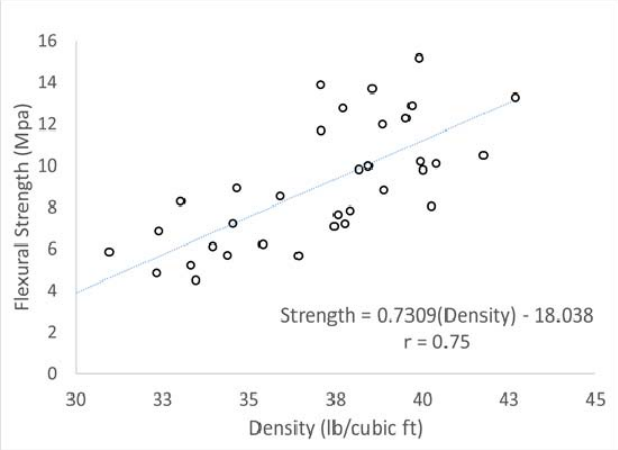

Fig. 1.
Composite strength depends on the density or mat weight (Via 2013).

However, the vast majority of manufacturers in the sustainable biomaterials industries and agricultural processing industries have not integrated the destructive testing and QC laboratories databases with the real-time process data warehouses. The industries typically have the destructive testing and QC laboratories databases on protected '*business networks*' and PLC data warehouses on separate protected '*process networks*.' Integrating large amounts of digital data stored in multiple databases across networks by data fusion and improving data quality is the gateway for discovery using '*Cloud Manufacturing*' and advanced data mining to support successful real-time predictive analytics.

METHODS

Automated Digital Data Integration and Data Fusion

An integrated database that fuses the process data with the QC lab destructive or product attribute test data was developed for users of the data depository. Users of the system will be required to upload databases for integration. Most databases in the agricultural industries at the mill-level are commercial where data are not encrypted (*e.g., Microsoft SQL, OSIsoft PI, Sytec IPS B.V.)*.[**] The databases when uploaded to the depository will be converted for SQL access where SQL scripting language will be used to accomplish the integration. After integration, the databases will be stored in SQL or comma separated variable (*.csv*) files which are easily convertible to almost every commercial software platform. The author has a breadth of experience with this task from prior published research (André *et al.* 2008, Clapp *et al.* 2008, Kim *et al.* 2012, Via 2013, André and Young 2013, Young *et al.* 2014, Young *et al.* 2015).

DATABASE INTEGRATION

In prior research studies (André *et al.* 2008, André and Young 2013, Young *et al.* 2014, Young *et al.* 2015), destructive QC lab results from several sustainable biomaterials manufacturers were matched with real-time process data from data warehouses using the *'date-time'* stamp of the destructive test sample selected from the production line. In prior studies, data were successfully combined into SQL tables that appear in a fused database to support additional real-time data mining and predictive analytical initiatives (Fig. 2).

[**] Microsoft SQL. 2018, https://www.microsoft.com/en-us/sql-server/sql-server-2017; OSIsoft PI. 2018, *https://www.osisoft.com/pi-system/;* Sytec IPS B.V. 2018, https://nld.bizdirlib.com/node/100715

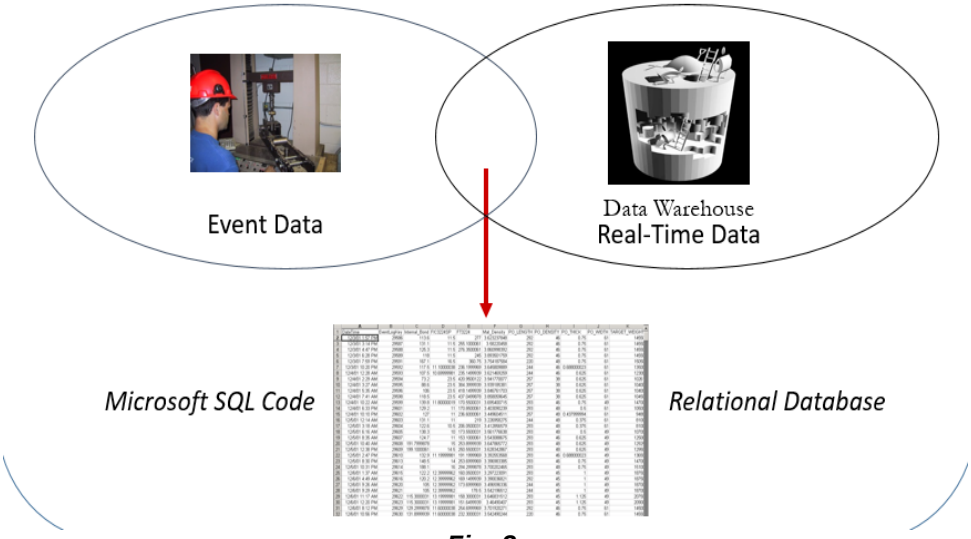

Fig. 2.
Illustration of combination of data warehouse (real time data) with destructive test laboratory data (event data).

The research in general will follow a similar approach where a Microsoft SQL table (*e.g., "InSQLPivot"*) will be created from the original databases that is a snapshot of the QC lab and process data. In prior research, an *"InSQLPivot_AV"* table contained all of the lab or QC data and median estimates of the last *n* = 100 records of real-time process data (*note, real-time process records typically are stored at rates of seconds and milliseconds*). The sampling record length (*n*) will be at the discretion of the user. An exclusive feature of the *"InSQLPivot_AV"* table was not only the automated alignment and fusion of real-time process data with lab data, but also the *'time-lagging'* of real-time process data which will also be part of the proposed system.

Data Fusion Process

The following methods were used in prior research and will vary depending on the number of databases and data structures that require fusion. For a two-database example, four tables are created within a Microsoft SQL database, for example:

- *InSQLPivot*
- *InSQLPivot_AV*
- *InSQLPivot_temp*
- *InSQLPivot_temp2.*

"InSQLPivot" stores the destructive lab and process data which consists of columns for both lab and process data. The field names for the lab data are preserved from the table where the lab data originated. The field names for the process data are identical to the tag names for the respective process parameters that are stored in these fields. In addition to these columns, a *"DateTime"* column that is used to record the time at which the process data were collected is essential. It is anticipated that sample *'tagnames'* associated with lab test time will vary by user of the system. The user will enter this necessary information in the user setup template. *"InSQLPivot_AV"* will consist of identical columns as the *"InSQLPivot"* table. *"InSQLPivot_temp"* and *"InSQLPivot_temp2"* are used for temporary storage of data while stored procedures are executing.

SQL Stored Procedures

In this example, four stored procedures are used to create the integrated database:

- *sp_Fill_InSQLPivot_From_Lab_DB*
- *sp_Fill_ InSQLPivot_AV_From_Lab_DB*
- *sp_Fill_From_eventsnapshot*
- *sp_Fill_From_summarydata.*

The stored procedure *"sp_Fill_InSQLPivot_From_Lab_DB"* records a copy of the QC lab data into the *"InSQLPivot"* table. The *"sp_Fill_ InSQLPivot_AV_From_Lab_DB"* stored procedure records a copy of the QC lab data into the *"InSQLPivot_AV"* table. The stored procedure *"sp_Fill_From_eventsnapshot"* records data from the *"v_eventsnapshot"* view (process database) into the *"InSQLPivot."* A stored procedure

"sp_Fill_From_eventsnapshot" records data from the "v_summarydata" view (generated from process database) into the *"InSQLPivot."* A *"v_summarydata"* data will match the lab data using a unique number (*e.g., idnum*).

SQL Data Transformation Services

In the aforementioned example, Data Transformation Services (DTS) are used to output the *"InSQLPivot"* and *"InSQLPivot_AV"* tables into two comma separated variable (.csv) files ("InSQLPivot.csv" and "InSQLPivot_AV.csv").

SQL Jobs

Three jobs are used to automate the process of inserting data into the *"InSQLPivot"* and *"InSQLPivot_AV"* tables and generate the final *.csv* files for the user to retrieve from the depositary, for example:

- *InSQLPivot_Filler*
- *InSQLPivot_AV_Filler*
- *Create_CSVs_From_InSQLPivots.*

The job *"InSQLPivot_Filler"* and *"InSQLPivot_AV_Filler"* are used to execute the stored procedures. The job *"Create_CSVs_From_InSQLPivots"* is used to execute the DTS Package *"Create_CSVs_From_InSQLPivots"* where the fused *.csv* files are generated (Fig. 3). The fused data files then undergo data quality assessment. The users will be notified by e-mail that the fused data files have been created and are undergoing data quality assessment.

Dynamic Time Lagging

A key and unique feature of the proposed research is dynamic time-lagging of process data. In all continuous agricultural and sustainable biomaterials processes, material flows past on-line sensors that measure attributes such as temperature, weight, moisture content, etc. This material passes the sensors at different times in the process relative to the '*end time*' at which the final product is manufactured, *i.e.,*

EventLogKey	DateTime	Comi	ProductI	TestNumber	ParallelEl	BunkerSpeed_BCL	BunkerSpeed_BSL	BunkerSpeed_TCL	BunkerSpeed_TSL	DividingRolls_BCL
26109	12/12/2005 19:53	ok	7/16 RS	100618	58047.475	41	69	41	69	164
26026	12/12/2005 14:03	ok	7/16 RS	100617	58653.225	39	61	39	61	164
25973	12/12/2005 8:05	ok	7/16 RS	100616	54024.95	42	72	42	72	164
25777	12/11/2005 19:45	ok	7/16 RS	100615	56764.7	46	73	44	73	164
25686	12/11/2005 14:18	ok	7/16 RS	100614	55526.075	43	74	43	74	164
25621	12/11/2005 7:30	ok	7/16 RS	100613	57665.1	43	69	43	70	164
25322	12/10/2005 14:35	ok	7/16 RS	100610	79498.925	39	67	40	67	164
25240	12/10/2005 8:27	ok	7/16 RS	100609	76640.95	44	71.5	44	72	164
24977	12/9/2005 19:19	ok	7/16 RS	100608	65828.8	41	68	41	68	164
24910	12/9/2005 13:53	ok	7/16 RS	100607	55315.675	44	61	44	61	164
24830	12/9/2005 8:29	ok	7/16 RS	100606	61387.775	48	64	47	64	164
24674	12/8/2005 22:16	ok	7/16 RS	100604	56287.475	37	60.5	37	61	164
24546	12/7/2005 19:35	ok	7/16 RS	100603	63515.325	43	66	43	66	164
24574	12/7/2005 14:36	ok	7/16 RS	100602	58926.3	44	71	44	71	164
24545	12/7/2005 9:09	ok	19/32 RS	100601	161153.475	39	63	39	63	164
24573	12/7/2005 2:26	ok	7/16 RS	100600	59603.325	29	48	29	48	164
24544	12/6/2005 19:11	ok	7/16 RS	100599	55381.3	39	66	39	66	164
24572	12/6/2005 14:18	ok	7/16 RS	100598	58865.25	46	66	46	66	164
24543	12/6/2005 8:39	ok	7/16 RS	100597	69422.25	39	57	39	57	164
24571	12/6/2005 0:31	ok	7/16 RS	100596	55201.25	41	68	41	68	164
24542	12/5/2005 20:17	ok	7/16 RS	100595	57121.55	41	66	41	66	164
24541	12/5/2005 7:17	ok	7/16 RS	100594	54550	29	45	29	45	164

Fig. 3.
Illustration of automated fused data file.

process data from sensors stored in the data warehouse are typically not aligned in time with material flow.

Users of the depository are given the option to enter time lags for the process sensor data (*e.g., tagnames*) stored in the process data warehouse. If the user does not have time lags, the user will be asked for distances (e.*g., inches, centimeters, feet, or meters*) from the end point of manufactured product and the location of the sensor associated with the *"tagname."* Distances are essential for dynamic time-lagging because of changes in the line speed of the process. This data fusion process may require time and motion, and material flow studies by the users at their project sites (Reigler *et al.* 2015). However, studying the time-lagging of sensors in the process relative to material flow is extremely beneficial for creating a high quality fused databases for data mining and real-time predictive analytics applications, see illustration in Fig. 4.

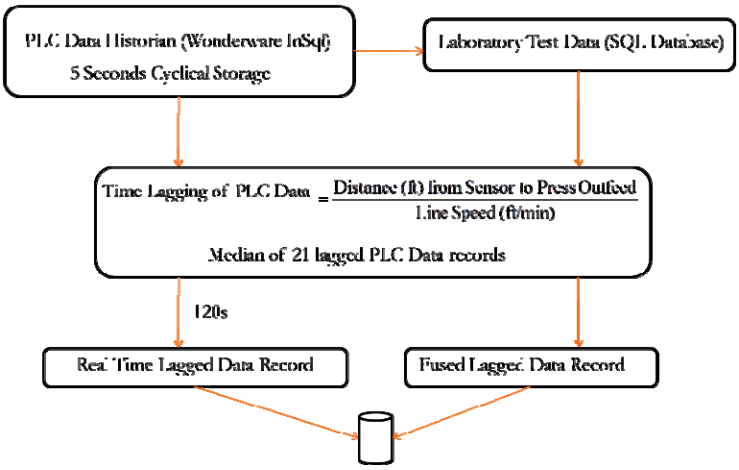

Fig. 4.
Flow chart illustration of data fusion process with examples of cycle time storage and lagged record retrieval.

DISCUSSION

The Importance of Data Quality

This research is aligned with the need for improved data quality from integrated databases. Francisco *et al.* (2017) and Ardagna *et al.* (2018) defined data quality (DQ) as an important issue for modern organizations, mainly for decision-making based on information. Francisco *et al.* (2017) indicated *"…that in order to obtain quality data, it is necessary to implement methods, processes, and specific techniques that handle information as a product, with well established, controlled, and managed production processes, e.g., TQDM – Total Quality Data Management."* Ardagna *et al.* (2018) stated *"…data can create a real value only if combined with quality: good decisions and actions are the results of correct, reliable and complete data, i.e., methods and techniques for the data quality assessment can support the identification of suitable data to process."* Hazen *et al.* (2014) remarked, *"…for today's supply chain professionals, there is an impetus for organizations to adopt and perfect data analytic functions (e.g., data science, predictive analytics, and big data) in order to enhance supply chain processes and, ultimately, performance. However, management decisions informed by the use of these data analytic methods are only as good as the data on which they are based.*" Cai and Zhu (2015) further discuss, "*high-quality data are the precondition for analyzing and using big data and for guaranteeing the value of the data,*" also see Batini *et al.* (2012) and Dumbill (2013). Gupta and Rani (2018) highlight the challenges in terms of "*data capture, storage, manipulation, management, analysis, etc., and the wide gap that exists between big data potential and realization*" given the many data quality shortcomings. Gupta and Rani (2018) indicated the need for research efforts in academia to assist industry in the understanding of big data and data quality. Liu *et al.* (2015) further articulate this point, "*big data brings lots of 'big errors' in data quality and data usage….information incompleteness is one of these problems*."

This research addresses the problem of information incompleteness (or information loss) and the development of automated algorithms to reduce information loss. However, research on '*data quality science*' to improve cloud manufacturing for the sustainable biomaterials and agricultural processing industries has not been well documented. Given this need, we propose to develop a data depositary and associated software systems for '*real-time automated data fusion*' and '*advanced algorithms for data quality improvement*,' both essential elements for successful data mining and predictive analytics.

Algorithms for Data Quality Assessment and Improvement

Outlier Assessment and Model Fitting. -- An outlier test will be performed on the datasets using the Grubbs' test, and others will be explored during the research. For example, the Grubbs' test calculates a test statistic iteratively for the smallest [1] and largest [2] values in the data set:

$$\text{Smallest value, } G = \frac{\bar{Y} - Y_1}{s} \quad [1]$$

$$\text{Largest value, } G = \frac{Y_n - \bar{Y}}{s} \quad [2]$$

where: $\bar{Y}$ is the sample mean,

Y_1 is the i[th] smallest value in the sample,
s is the standard deviation,
n is the number of observations in the sample.

Probability density functions (*pdfs*) for the variables in the datasets will be assessed using the Aikake's Information Criteria (AIC) and Bayesian Information Criteria (BIC) test statistics (Akaike 1974, Schwarz 1978). The AIC [3] and BIC [4] are given by:

$$AIC = 2k - 2\ln(\hat{L}) \quad [3]$$
$$BIC = \ln(n)\,k - 2\ln(\hat{L}) \quad [4]$$

where:
$\hat{L}$ is the maximized value of the likelihood function of the model,
k is the number of parameters estimated by the model,
n is the number of observations in the sample.

AIC and BIC test statistics will be generated for the following probability density functions (pdfs): Normal, Gamma, Lognormal, Weibull, Logistic, Loglogistic, Largest Extreme Value, Exponential, and Frechet. A report will be generated for users indicating which data points are outliers for the variables (tagnames) of the datasets and which *pdfs* are the best fits for each variable or tagname. Generating a list of best fit for *pdfs* may be important for the dependent variables (*Y*), or data in the QC lab or destructive test datasets. Many data mining applications of a stochastic nature have important assumptions for the *pdfs* of the dependent variables.

Imputation Algorithms for Missing Fields. -- The authors have strong knowledge of data imputation for process data from manufacturing processes. In the study by Zeng *et al.* (2016), maximum-likelihood based imputation using the expectation-maximization (EM) algorithm and multiple imputation (MI) method with Markov Chain Monte Carlo (MCMC) simulation were used to reduce information loss for improved predictive analytics and data mining (Fig. 5). Other methods of imputation will also be tested. The mean/median substitution method replaces missing fields with the mean/median of the corresponding variable (Little 1992). The last observation carried forward (LOCF) method simply replaces missing values with the last known value of the variable in a time-ordered data set, see Gelman and Hill (2007), and Hamer *et al.* (2009). The simple random imputation method ('*hot-deck method*') replaces missing values with a randomly selected value from another observation of the same variable (Altmayer 2002, Lanning and Berry 2003).

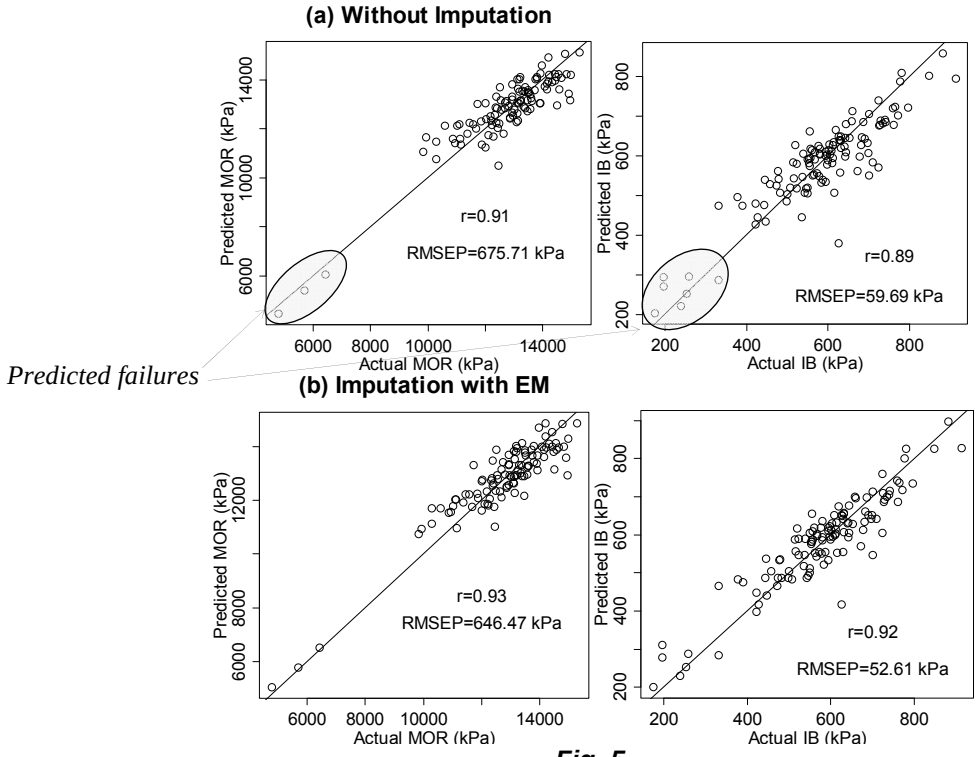

Fig. 5.
XY scatter plots and correlation coefficients for validation data set predicting MOR and IB using Bayesian Additive Regression Tree (BART) models for a) top two graphs without imputation; b) bottom two graphs with imputation (Zeng et al. 2016).

In general, the EM algorithm works as follows: 1) fill-in the missing data, Y_{mis}, based on $\theta^{(0)}$, 2) re-estimate θ based on the now complete data, 3) and iterate until convergence. As Dempster *et al.* (1997) observed, EM does not directly estimate the missing data, just the functions of missing data that appear in the likelihood function, which in many cases are much simpler to estimate than the entire missing dataset (Agarwal 2001). MI imputation methods replace each missing value with a set of plausible values that represent the uncertainty of the true value. The key steps can be summarized as follows (Allison 2000): 1) impute missing values using an appropriate model such as regression model that incorporates random variation; 2) repeat this process *N* times, producing *N* complete data sets; 3) perform the desired analysis on each data set using standard complete-data methods; 4) average the values of the parameter estimates across the *N* samples to produce a single point estimate. A direct approach for MI is MCMC (Schafer 1997). In general, MCMC is a collection of techniques for obtaining pseudorandom draws from a probability distribution.

In Zeng *et al.* (2016) the EM and MI algorithms when compared with mean/median substitution, last-value-carried-forward (LOCF), and *'hot-deck'* random replacement, provided better results in prediction for the test data for final quality attributes of manufactured product. In general Zeng *et al.* (2016) found that better predictive analytics (and predicting failures in a process) were achieved from imputed datasets than predictive analytics based on non-imputed datasets. A substantial achievement lending credibility to this proposal is that the imputed data sets were tested in an industrial setting where real-time data were collected and the aforementioned algorithms were used to improve the quality of a manufacturers fused data sets. All of the aforementioned algorithms will be developed as one system for reducing information loss. Real-time data validation tests will be part of the system where the best imputation methods will be selected for the use in the final database. However, the user of the system may ask to have separate imputed databases by method which will developed for them.

Depositary Database and Software Platform

The key steps in the system are that the participants of the data depositary compete key information about the data structure and '*tagnames*' that will link the digital data. A secure VPN portal will be established and participants will be sent a three-level password to enter the portal. All data will be encrypted upon entry into the portal and also encrypted upon exit of the portal. After data fusion and data improvement a report will be generated for all users that identify key attributes of the fused data and the data quality improvement that occurred, *e.g., number of outliers removed by tagname, records removed, level of data imputation, and probability density function by tagname.* The proposed data depository is illustrated in Fig. 6.

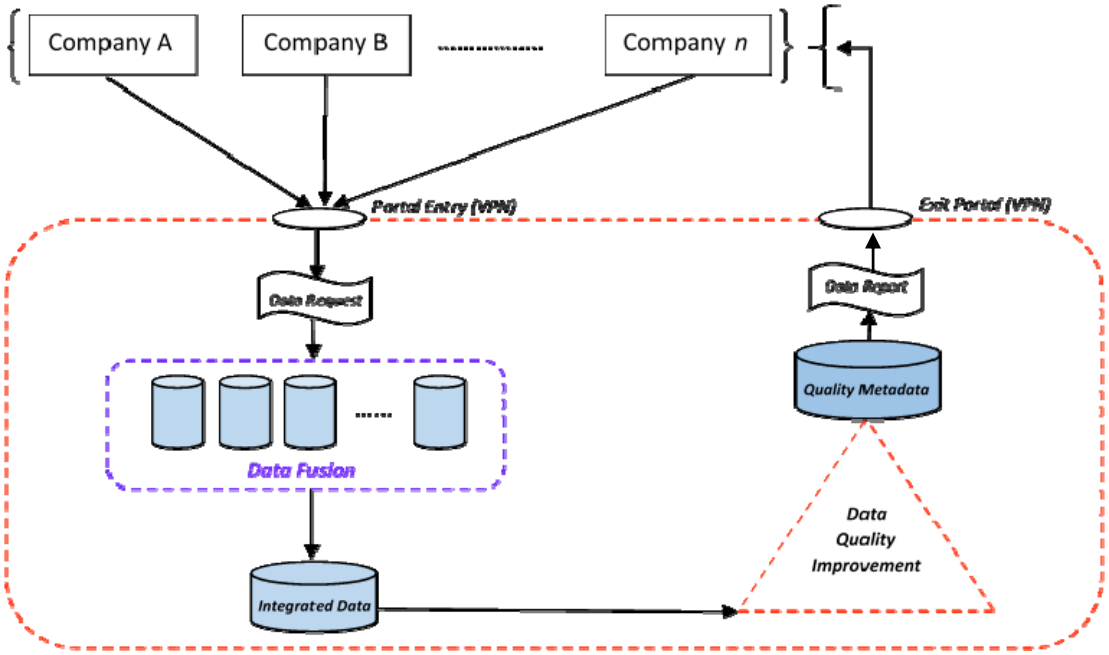

Fig. 6.
Illustration of data depository with data fusion and data quality improvement.

REFERENCES

Akaike H (1974) A new look at the statistical model identification. IEEE Transactions on Automatic Control. 19(6):716–723. https://doi.org/10.1109/TAC.1974.1100705

Agarwal S (2001) Learning from incomplete data. Department of Computer Science and Engineering, University of California, San Diego. http://cseweb.ucsd.edu/~elkan/254spring01/sagarwalrep.pdf

Allison P (2000) Multiple imputation for missing data: a cautionary tale. Sociological Methods and Research. 28:301-309.

Altmayer L (2002) Hot-deck imputation: a simple data step approach. *Proc.* 2002 Northeast SAS User's Group. Buffalo, NY: Northeast SAS User's Group. pp. 773–780.

American Forest and Paper Association (2018) Economic Impact of the forest products industry. On-line publication. http://www.afandpa.org/our-industry/economic-impact Last accessed on 16 August 2018.

André N, Young TM (2013) Real-time process modeling of particleboard manufacture using variable selection and regression methods ensemble. European Journal of Wood and Wood Products. (Eur. J. Wood Prod. Holz als Roh- und Werkstoff). 71(3):361-370. https://doi.org/10.1007/s00107-013-0689-0

André N, Cho, Baek SH, Jeong MK, Young TM (2008) Prediction of internal bond strength in a medium density fiberboard process using multivariate statistical methods and variable selection. Wood Science and Technology 42(7):521-534. https://doi.org/10.1007/s00226-008-0204-7

Ardagna D, Cappiello C, Samá W, Vitali M (2018) Context-aware data quality assessment for big data. Future Generation Computer Systems. 89:548–562. https://doi.org/10.1016/j.future.2018.07.014

Atta-Obeng E, Via BK, Fasina O (2012) Effect of microcrystalline cellulose, species, and particle size on mechanical and physical properties of particleboard. Wood and Fiber Science. 44(2):227-235.

Barton D, Court D (2012) Making advanced analytics work for you. Harvard Business Review. 90(10):78–83,128.

Batini C, Scannapieco M (2006) Data quality: concepts, methodologies and techniques. Berlin Heidelberg: Springer-Verlag.

Blanc P, Demongodin I, Castagna P (2008) A holonic approach for manufacturing execution system design: An industrial application. Engineering Applications of Artificial Intelligence, 21(3):315–330. https://doi.org/10.1016/j.engappai.2008.01.007

Cai L, Zhu Y (2015) The challenges of data quality and data quality assessment in the big data era. Data Science Journal. 14(2):1-10. https://doi.org/10.5334/dsj-2015-002

Carty DM, Young TM, Zaretzki RL, Guess FM, Petutschnigg A (2015) Predicting the strength properties of wood composites using boosted regression trees. Forest Products Journal. 65(7/8):365-371. https://doi.org/10.13073/FPJ-D-12-00085

Chee CH, William Y, Shijia G, Richards G (2014) Improving business intelligence traceability and accountability. Journal of Database Management. 25(3):28–47. https://doi.org/10.4018/jdm.2014070102

Clapp NEJr., Young TM, Guess FM (2008) Predictive modeling the internal bond of medium density fiberboard using a modified principal component analysis. Forest Products Journal. 58(4):49-55.

Deepak G, Rani R (2018) A study of big data evolution and research challenges. Journal of Information Science. https://doi.org/10.1177/0165551518789880

Deming WE (1986) Out of the crisis. Massachusetts Institute of Technology, Center for Advanced Engineering Study. Cambridge, MA. 507p.

Deming WE (1993) The new economics. Massachusetts Institute of Technology, Center for Advanced Engineering Study. Cambridge, MA. 240p.

Dempster AP, Laird NM, Rubin DB (1977) Maximum likelihood from incomplete data via the EM algorithm. Journal of the Royal Statistical Society, Series B. 39(1):1-38.

Dumbill E (2013) Making sense of big data. Big Data. 1(1):1–2. https://doi.org/10.1089/big.2012.1503

Dunham and Associates (2017) U.S. Food and Agriculture Industries Economic Impact Study, 2017. Brooklyn, NY. http://www.feedingtheeconomy.com/assets/res/methodology.pdf

Francisco MMC, Alves-Souza SN, Campos EGL, De Souza LS (2017) Total data quality management and total information quality management applied to customer relationship management. *Proc.* ICIME 2017, October 9–11, 2017, Barcelona, Spain. https://doi.org/10.1145/3149572.3149575

Fosso Wamba S, Akter S, Edwards A, Chopin G, Gnanzou D (2015) How “big data” can make big impact: Findings from a systematic review and a longitudinal case study. International Journal of Production Economics. 165:234–46.

Gandomi A, Haider M (2015) Beyond the hype: Big data concepts, methods, and analytics. International Journal of Information Management. 35:137–144. https://doi.org/10.1016/j.ijinfomgt.2014.10.007

Gelman A, Hill J (2007) Data analysis using regression and multilevel/hierarchical models. Cambridge University Press, Cambridge.

Grubbs FE (1950) Sample criteria for testing outlying observations. Annals of Mathematical Statistics. 21(1):27–58. https://doi.org/10.1214/aoms/11777298855

Hamer RM (2009) Last observation carried forward versus mixed models in the analysis of psychiatric clinical trials. American Journal of Psychiatry. 166:639-641.

Hazen BT, Boone CA, Ezell JD, Jones-Farmer LA (2014) Data quality for data science, predictive analytics, and big data in supply chain management: An introduction to the problem and suggestions for research and applications. International Journal of Production Economics. (154):72-80. https://doi.org/10.1016/j.ijpe.2014.04.018

Kim H, Guess FM, Young TM (2011) An extension of regression trees to generate better predictive models. IIE Transactions. 43(1):43-54. https://doi.org/10.1080/0740817X.2011.590441

Kim N, Jeong YS, Jeong MK, Young TM (2012) Kernel ridge regression with lagged dependent variable: applications to prediction of internal bond strength in a medium density fiberboard process.

IEEE Transactions on Systems, Man, and Cybernetics - Part C: Applications and Reviews. 42(6):1011-1020. https://doi.org/10.1109/TSMCC.2011.2177969

Lanning D, Berry D (2003) An alternative to PROC MI for large samples. SAS Users Group International (SUGI) 28, Seattle, Washington.

Lasi H, Fettke P, Kemper HG, Feld T, Hoffmann M (2014) Industry 4.0. Business Informtaion Systems and Engineering. 6(4):239–42.

Lee J, Kao HA, Yang S (2014) Service innovation and smart analytics for Industry 4.0 and big data environment. Procedia CIRP 16:3–8.

Lee J, Lapira E, Bagheri B, Kao H (2013) Recent advances and trends in predictive manufacturing systems in big data environment. Manufacturing Letters. 1(1):38–41.

Liao Y, Deschamps F, de Freitas Rocha Loures E, Pierin Ramos LF (2017) Past, present and future of Industry 4.0 – a systematic literature review and research agenda proposal. International Journal of Production Research. 55(12):3609–3629. https://doi.org/10.1080/00207543.2017.1308576

Little RJA (1992) Regression with missing X's: a review. Journal of the American Statistical Association. 87:1227-1237.

Manyika J, Chui M, Brown B, Bughin J, Dobbs R, Roxburgh C (2011) Big data: The next frontier for innovation, competition, and productivity. New York: McKinsey Global Institute.

Redman T (1996) Data quality for the information age. Norwood: Artech House.

Rich S (2012) Big data is a “new natural resource,” IBM says. Jun 27 Available from: http://www.govtech.com/policy-management/Big-Data-Is-a-New-Natural-Resource-IBM-Says.html accessed on 13 August 2018.

Riegler M, André N, Gronalt M, Young T (2015) Dynamic simulation of the continuous flow of bulk material during production to improve the statistical modeling of final product strength properties. International Journal of Production Research. 53(21):6629–6636. https://doi.org/10.1080/00207543.2015.1055844

Rose E (1991) Data modeling for non-standard data. Journal of Database Management. 2(3):8–21. https://doi.org/10.4018/jdm.1991070102

Schafer JL (1997) Analysis of incomplete multivariate data. Chapman & Hall, London.

Schwarz GE (1978) Estimating the dimension of a model. Annals of Statistics. 6(2):461–464. https://doi.org/10.1214/aos/1176344136

Taguchi G, Chowdhury S, WU Y (2005) Taguchi's quality engineering handbook, Hoboken, New Jersey, John Wiley & Sons, Inc.

Taylor A, Via BK (2009) Potential of visible and near infrared spectroscopy to quantify phenol formaldehyde resin content in oriented strandboard. European Journal of Wood and Wood. Products. 67(1):3-5.

Theorin A, Bengtsson K, Provost J, Lieder M, Johnsson C, Lundholm T, Lennartson B (2015) An event-driven manufacturing information system architecture for Industry 4.0. International Journal of Production Research. 55(5):1297–1311. https://doi.org/10.1016/j.ifacol.2015.06.138

Thomas B, Jacob G (2016) Big data – the change of datacenter concept and challenges. International Advanced Research Journal in Science, Engineering and Technology. 3(7):261-263. https://doi.org/10.17148/IARJSET.2016.3754.

Tian N, Young TM, Petutschnigg A, Sun Y, Pei Z (2018) Improved predictive modeling using Bayesian Additive Regression Trees (BART) for wood composite products. International Journal of Engineering and Science (IJSE). 7(5):66-75. https://doi.org/10.9790/1813-0705036675

U.S. Department of Agriculture (2018) Employment opportunities for college graduates in food, agriculture, renewable natural resources, and the environment in the United States, 2015-2020. https://www.purdue.edu/usda/employment/. Last accessed on 16 August 2018.

U.S. Department of Energy (2018) The forest products industry. Office of Energy Efficiency and Renewable Energy. U.S. Department of Energy 2018. On-line publication. https://www.energy.gov/eere/amo/forest-products. Last accessed on 16 August 2018.

Via B (2013) Characterization and evaluation of wood strand composite load capacity with near infrared spectroscopy. Materials and structures. 46(11):1801-1810.

Via B, McDonald T, Fulton J (2012) Nonlinear multivariate modeling of strand density from near-infrared spectra. Wood Science and Technology, 46(6):1073-1084.

Via BK, Hand WG, Banerjee S (2016) U.S. Patent Application No. 15/067,729.

Wang S, Wan J, Zhang D, Li D, Zhang C (2016) Towards smart factory for Industry 4.0: A self-organized multi-agent system with big data based feedback and coordination. Computer Networks. 101:158–68.

Wang SY, Wan J, Li D, Zhang C (2016) Implementing smart factory of Industrie 4.0: An outlook. International Journal of Distributed Sensor Networks. 2016:3159805.

Young T, André N, Otjen J (2015) Quantifying the natural variation of formaldehyde emissions for wood composite panels. Forest Products Journal. 65(3/4):S82-S84.

Young TM, Clapp NEJr, Guess FM, Chen CH (2014) Predicting key reliability response with limited response data. Quality Engineering. 26(2):223-232 https://doi.org/10.1080/08982112.2013.807930

Young, TM, Shaffer LB, Guess FM, Bensmail H, León RV (2008) A comparison of multiple linear regression and quantile regression for modeling the internal bond of medium density fiberboard. Forest Products Journal. 58(4):39-48.

Young TM, León RV, Chen CH, Chen W, Guess FM, Edwards DJ (2015) Robustly estimating lower percentiles when observations are costly. Quality Engineering. 27:361-373. https://doi.org/10.1080/08982112.2014.968667

Zeng Y, Young TM, Edwards DJ, Guess FM, Chen CH (2016) Case studies: A study of missing data imputation in predictive modeling of a wood composite manufacturing process. Journal of Quality Technology. 48(3):284-296. https://doi.org/10.1080/00224065.2016.11918179

Zhong RY, Xun X, Klotz E, Newman ST (2017) Intelligent manufacturing in the context of Industry 4.0: a review. Engineering 3:616–630. https://doi.org/10.1016/J.ENG.2017.05.015

Zhong RY, Newman ST, Huang GQ, Lan S (2016) Big data for supply chain management in the service and manufacturing sectors: Challenges, opportunities, and future perspectives. Computers and Industrial Engineering. 101:572–591.