J Pharm Innov (2007) 2:6–17
DOI 10.1007/s12247-007-9005-z

Capitalizing on Aggregate Data for Gaining Process Understanding—Effect of Raw Material, Environmental and Process Conditions on the Dissolution Rate of a Sustained Release Product

Karel Stryczek · Petr Horacek · Jiri Klema · Xavier Castells · Bob Stewart · Jean-Marie Geoffroy

Published online: 17 October 2007

Abstract Continuous improvement of pharmaceutical manufacturing operations has not evolved at the same rate as it has in other industries. Although time-series data are routinely collected as part of equipment control systems, the data are usually not thoroughly evaluated. This article investigates batch data, in-process and release laboratory test data and time-series data from granulation, fluid-bed drying and coating operations in an effort to determine which parameters are most critical to the dissolution of a matrix-release, solid oral dosage form of a poorly soluble drug.

Keywords Time-series data · Critical process parameters · Correlation-based feature selection · Quality by design · Pharmaceutical process modeling · Critical quality attributes (CQAs)

K. Stryczek
Rockwell Automation, 1 Allen-Bradley Drive, Mayfield Heights, OH 44124, USA

P. Horacek
Rockwell Automation, Research Center Prague, Pekarska 695, Prague 5, 15500, Czech Republic

J. Klema
Gerstner Laboratory for Intelligent Decision Making and Control, Department of Cybernetics, Czech Technical University, Technicka 2, Prague 6, 16627, Czech Republic

X. Castells · B. Stewart
Abbott Laboratories, 1401 N Sheridan Rd, N. Chicago, IL 60064, USA

J.-M. Geoffroy (✉)
TAP Pharmaceuticals, Inc., 675 N. Field Drive, 0T85, Bannockburn, Lake Forest, IL 60045, USA
e-mail: jean-marie.geoffroy@TAP.com

Introduction

The scope of this study was to analyze historical process data associated with a modified-release, solid oral tablet manufacturing process. Production data from 211 lots of drug product, covering more than 140 parameters were collected. The goals of the study were to identify critical process parameters (CPPs) from the historical data and to determine which modelling technique best applies for this type of dosage form.

Several automatic parameter selection and ranking methods were applied in order to obtain a list of CPPs influencing the tablet dissolution rates. During consequent modeling, the influence of CPPs was thoroughly analyzed first, and then used to predict the dissolution rates; the methods included empirical models, regression tree models and autoregressive models.

The correlation-based feature selection (CFS) method, used for CPP ranking, yielded results most consistent with the existing engineering understanding of the process. This method was identified as the best for automatic ranking of parameters based on their respective impact on dissolution rates. The presented results demonstrate that the long-term variation in the dissolution rates for a modified-release tablet are caused by granulation process conditions, external ambient conditions and raw material properties. The predictive mathematical models that were used provided accurate and reliable predictions of dissolution, based on data available early in the production cycle.

Understanding the impact of individual process parameter variation on critical product properties, such as dissolution rate, leads to the following opportunities for process understanding and control:

- Process modeling and predictions of key product properties;

- Real-time control of key product properties;
- Optimization of manufacturing processes.

Control of dissolution rates for a sustained-release product can often be challenging. Numerous articles have been published on traditional process variables that affect the dissolution performance of immediate and sustained-release dosage forms [1–12]. The aim of this work was to determine if a deeper understanding of dissolution might be possible by analyzing all available raw material properties, process parameters and external environmental conditions. In addition, it is desirable to know which attributes are worthy of further study, either through tighter monitoring or through active experimentation with statistical experimental designs.

Examples of investigated attributes include:

- Temperature and humidity conditions outside a facility that are both temperature and humidity controlled;
- Impurities in excipients and drug substance;
- Process parameters such as granulation power and time, and mean and maximum drying temperatures, tablet coating conditions including temperature and airflow rates;
- Traditional process parameters such as tablet hardness and drug substance particle size.

Experimental

Description of Manufacturing Process

The formulation consists of drug, sustained-release polymer and other standard manufacturing ingredients. The manufacturing process consists of high-shear, wet granulation, fluid-bed drying, milling, final blending with all remaining ingredients, including magnesium stearate, and compressing and coating. The coating is non-functional.

Summary of Process Variables Investigated

Historical data for more than 140 raw materials, environmental and process parameters were obtained from 211 lots of product. This includes 1,688 granulation sub-runs (8 per lot), 1,055 coating sub-runs (5 per lot) and ~8,000 dissolution tablet tests.

If more than one lot of a raw material was used in a batch during manufacture, raw material attributes were calculated by taking the weighted average from each lot of each raw material property, as calculated from the bill of materials for that lot and raw material test data.

Process parameters, in-process testing and release testing data were obtained (and if necessary, calculated) from the electronic data warehouse.

External air temperature, dew points, relative humidity and precipitation data were obtained from a local National Weather Service station.

Interviews with subject matter experts were conducted in the production facility. The purpose of these interviews was:

- To obtain a detailed description of the unit operations, equipment, and operator interfaces;
- To identify the ‘obvious’ information, routine operational issues;
- To collect important information from operators with practical experience in running the process;
- To collect additional detail from process experts (expert knowledge represents additional data in the form of, for example, ‘if...then...else’ rules);
- To identify key points of interest and expectations from individual stakeholders.

See Table 1 for a partial list of analyzed raw material parameters and process conditions, and Table 2 for parameter counts.

Data Pre-Processing—Purpose and Methodology

The objective of pre-processing is to bring together the data representing different manufacturing phases to analyze and model potential relationships. During pre-processing, data from various sources and formats is converted into one electronic data file in a standard format so that it is possible to identify the CPP using mathematical software tools.

The data sources used for this analysis consisted of batch processing data (from electronic batch records), laboratory test data (in-process and release testing), time-series data obtained from the facilities data historian, and information from subject matter experts.

Normalization of data was accomplished by taking the weighted-average for each raw material attribute or process parameter for each batch and dividing by its standard deviation. In this way, it is possible to estimate the effect across many variables. Moving-average filtering was used to reduce the effect of random measurement errors on the CPP analysis.

The first step in data pre-processing is to convert data from the individual sources, such as MS Excel, text files and paper records, into database tables. MS ACCESS was used to generate the tables containing the individual manufacturing phases and weather data and relate them to the dissolution data. The goal of this step is to understand the manufacturing process and identify ‘obvious’ CPP with respect to drug release in terms of single-dimensional analysis. This process is often referred to as pre-identification. The data can be aggregated either by SubRuns (granulation or coating SubRuns) or ProdRuns (individual production lots).

Table 1 List and description of selected parameters referenced in the analysis

Parameter	Units	Description
Release rate parameters		
R30	%	Drug release (%) after 30 min
R45	%	Drug release (%) after 45 min
R60	%	Drug release (%) after 60 min
Selected raw material, process and environmental parameters		
SRA_X	°C	Sustained releasing agent material property
Outside_Air_Temp	°C	Ambient air temperature (local airport)
Outside_Air_Dew_Point	°C	Ambient air dew point (local airport)
Outside_Air_Pressure	PSI	Atmospheric air pressure (local airport)
Precipitation	Inch	Daily average precipitation (local airport)
Water_Addition	Liters	Total water added per production run during granulation
IngrBulkVolume	Liters	Bulk volume of one standard manufacturing ingredient
AvgOfGPower_max	kW	Average of maximum power applied during granulation, per lot
TDiff_disp_release	^{0}C	Maximum minus minimum temperature differential over period of time elapsed between raw material dispensing and laboratory testing for dissolution (per lot)
API_repr	N/A	Drug substance property

130 other parameters were analyzed with respect to release rates. These parameters were identified as less significant during the CPP ranking process and were not listed in this table. The unlisted parameters were not identified as critical.

Second, all tables are joined using structured query language (SQL) to generate a source database. This table (database or matrix) aggregates the data by individual lots with the data available because this is the only common unit that appears in all the manufacturing steps.

Next, the resulting tables are exported into an arbitrary analytical and modeling software package such as MS Excel, MATLAB or WEKA. MS Excel, a well-known spreadsheet tool, was used for basic data manipulation and data sharing. The MATLAB product family provides a high-level programming language, an interactive technical computing environment and provides algorithm development, data analysis and/or visualization and numeric computation. WEKA [13] is a collection of machine learning algorithms for data mining tasks implemented in Java; WEKA contains tools for data pre-processing, classification, regression, clustering, association rules and visualization.

The global data analysis focuses on numerical parameters. The data regarding operators is challenging to analyze because each lot (ProdRun) is affected by the number (typically more than ten) of operators. The operator ID data has been analyzed in terms of single-dimensional analysis for each unit operation. The analysis has not shown dependence of release rates on the actions of any specific operator, groups of operators or shifts of operators.

The parameters that do not change within the analyzed time period have been removed.

Table 2 Summary of process parameter counts for raw materials, environmental process and other conditions

Unit operation	Number of variables
Dispensing	35
Granulation	12
Drying	8
Blending	9
Compressing	14
Coating and coating prep.	15
Other parameters	
Weather	25
Hold times (lot time duration)	5
Release	9
Miscellaneous	9
Total number of variables	141

Methods and Tools

The main goal of the study is to better understand the manufacturing process and consequently improve the control of critical product properties. This goal is broken down into two consecutive steps: CPP identification and understanding of the production operations; and quantification of how the analyzed process parameters influence the target product properties.

In terms of general methods, we speak about feature selection and modeling. The following section describes the methods that were used.

Identification of CPP—Feature Selection

CPP identification can be defined as a problem of ‘feature selection’, one of the central issues in machine learning or statistics. The main goal of CPP identification is to find a

set of parameters that have a strong influence on a target variable (e.g. dissolution, uniformity or yield). The feature selection procedure reduces a set of features to eliminate redundant, irrelevant or noisy features that do not help to increase classification or the prediction accuracy of a constructed model. The models strive to classify or predict the target variables (e.g. dissolution rates), whereby preliminary feature selection results in better model performance and reduced computation. The approach used in the CPP identification effort is described as follows: first, attempt to construct a model that best fits the target variable using a selected feature selection technique; second, a CPP becomes a feature, selected to be influential with respect to the model performance.

Feature selection techniques can be categorized according to several criteria. Bias criterion refers to whether the learning bias is guided by feedback from the learning algorithm performance or whether it is, instead, a preset bias that uses general characteristics of the data and operates independently of any learning algorithm. The first method is referred to as the 'wrapper' approach; the second method is often being referred to as the 'filter' approach [14]. A different taxonomy divides algorithms into those that evaluate and rank individual features and those that evaluate subsets of features. Many feature selection techniques handle regression problems, that is they deal with numeric target variables. The target variables in our domain are entirely numeric and, therefore, we focused on these techniques.

The simplest way to identify a CPP is to apply a filter approach using correlation as the underlying feature score function. In this approach, correlations between the target variable and all the parameters (features) are calculated; then, the features with the strongest correlation to the target variable are assumed to be CPP. The disadvantage is that it considers linear dependence only and does not account for mutual interactions between the investigated features. For example, the data suggest that weather conditions strongly influence dissolution; however, when another parameter is also influenced by weather it can easily appear that this parameter is also critical when it might have no actual connection to release. This correlation is often referred to as spurious. It should be noted that the primary cause might also be missing in the collected set of parameters.

Therefore, it is better to evaluate all subsets of features. Correlation-based feature selection (CFS) applies subset evaluation heuristics [14] (M.A. Hall, PhD thesis, Waikato University, 1998) and takes into account both the usefulness of the individual features for predicting the target variable and the level of intercorrelation between them. The heuristic approach prefers subsets that tend to correlate highly with the target variable despite having low intercorrelation between the individual features. Similarly, it is possible to use a linear learning algorithm within the wrapper framework.

A specific implementation of this approach is used in this project, and can be simplified in the following steps:

(1) Build and test all single variable linear models;
(2) Select the best model, that is the model that best fits the target function;
(3) If the model meets the target condition, then stop—the features in the model are CPPs and the target condition has been met;
(4) Attempt to add another feature into the best model, testing all features that are outside the model;
(5) Go to step (2).

The target conditions can vary according to the needs of the user. The first typical condition is the number of features used in the best model—the user determines the number of features to be selected a priori. The second typical condition is based on model performance—the features are added until the target performance is reached or until it improves sufficiently.

The following example demonstrates model development according to the previous implementation protocol. First, consider Release30 as the target function and begin by searching for two (2) CPPs out of one hundred and forty (140) features:

(1) $R30 = x11 \times API\_repr + x12$,
(2) $R30 = x21 \times Outside\_Air\_Temp + x22$, etc.

where $x11$, $x12$, $x21$ and $x22$ are constants optimized by regression analysis; API_repr is a drug substance property and Outside_Air_Temp is the average temperature during lot processing.

These parameters are 2 of 140 analyzed features; the number of models corresponds to the number of features, that is 140 models are built in this step.

Mean squared error is then used as the selection criterion. Assuming that model (2) is the best.

Outside_Air_Temp is the first CPP; therefore, the search for the second CPP can be initiated:

(1) $R30 = y11 \times Outside\_Air\_Temp + y12 \times API\_repr + y13$,
(2) $R30 = y21 \times Outside\_Air\_Temp + y22 \times Hardness + y23$, etc.

The number of models corresponds to the number of features, in this case 139 models are built in this step. Outside_Air_Temp appears in all the models.

Assuming model (2) be the best, the two most important CPPs have been identified: Outside_Air_Temp and (tablet) Hardness.

The previous example illustrates the use of the algorithm with two actual parameters. A detailed overview of identified CPP can be found in the "Results and Discussion" section.

Although the algorithms mentioned above can deal with the feature dependence, they are limited in that they are linear. Within this CPP identification effort, the authors analyzed parameters that were controlled during the manufacturing process and, therefore, only small perturbations appear in the majority of the features. It can be assumed that the linear approximation can fit the data dependencies well within the provided ranges of variation. Nevertheless, the dependencies do not have to be always truly linear.

The RReliefF [15] method provides an algorithm for dealing with nonlinearities that stems from the filter-based Relief algorithm. The key idea of the original Relief algorithm [16] is to estimate the quality of features according to how well their values distinguish between the instances that are near to each other. For that purpose, given a randomly selected instance (R), Relief searches for its two nearest neighbors: one from the same class, called nearest hit H, and the other from a different class, called nearest miss M. It updates the quality estimation for all the features depending on the values for R, M, and H. The goal is to minimize R–H distance and maximize its R–M counterpart. The process is repeated m times, where m is a parameter of the method.

Relief cannot be applied to regression tasks directly. Its essential disadvantage is the need for discretization of both continuous features and the target function. Consequently, all values within a given interval are treated as equal, which will result in a loss of information. RReliefF is a modification of the original algorithm that overcomes this bottleneck [16]: R refers to regression domains, and the final (F) stands for use of more hits and misses (neighbors) for each instance. RReliefF generates a feature ranking. In the CPP identification effort, this model is run more times with various random sampling of the original set of instances. The final output is then an average rank of each feature, and CPPs are those features with the lowest average rank. WEKA implementation of this approach, as used in this analysis, can be simplified as follows:

(a) Initiate a set of feature importance weights to zero;
(b) Repeat for m times (m is an algorithm parameter): randomly select an instance R_i (i.e. a ProdRun);
(c) Select k instances I_j, j=1. k, nearest to R_i (i.e. find such ProdRuns whose feature description best agrees with R_i).

For all the features to be assessed do the following:

(d) Repeat for all I_j and study whether the feature value in R_i and I_j changes in accordance with changes of the target value;
(e) If so, increase the feature importance weight (and vice versa);
(f) Use the feature importance weights to generate a feature ranking.

Mutual information (MI) [17] is an alternative that can replace the correlation measure. MI is a natural measure of the dependence between random variables and it is equivalent to the well-known Kullback–Leibler divergence between the joint density and the product of its marginal densities, a natural measure for independence. It is always non-negative and zero if, and only if, the variables are statistically independent. Thus, MI takes into account the whole dependence structure of the variables and not only the covariance, as does principal component analysis (PCA) and related methods. MI can also be used in the CFS algorithm, where correlation and intercorrelations are replaced by the measure of mutual information.

Modeling

In general, modeling follows two objectives—prediction and description [18]. The predictive models are used to predict future values of a target variable, their ultimate goal being the accuracy of the prediction. The descriptive models capture the relationships among variables and serve mainly to explain relationships among the target variable and parameters (independent variables). In the CPP project, both accurate prediction and description were two separate objectives. The following subsections discuss several different ways to acquire descriptive and predictive models and to model evaluation methods.

Descriptive Models

One of the main goals in manufacturing is to control actively and accurately the target variables. To achieve such control, the influence of the individual parameters on the target variable must be thoroughly understood. This is the main reason why we emphasize the need for simple and, thus, understandable descriptive models that deal with a few parameters only. The selected parameters used for modeling result from the feature selection procedures already described.

Initially, the linear models with empirically acquired constants were used to describe relationships demonstrated in the historical datasets. These models use limited sets of the top-ranked CPP for modeling. The main goal of building the models is to understand to what extent the most critical parameters influence the target variables. In this text we refer to this class of models as empirical models.

A regression tree model might be considered a variant of decision trees [19], designed to approximate real-valued functions instead of being used for classification tasks. Regression trees differ from decision trees mainly in having values rather than class labels at the leaves. Another variant of regression trees, also referred to as model trees [20], builds multivariable linear models at the leaves. These model trees

are thus analogous to piecewise linear functions. The model trees can deal with both discrete and continuous parameters, first separating the input space into characteristic regions and then building independent linear models representing the given regions. Consequently, the resulting tree can identify and truly approximate potentially complex, non-linear relationships. Complex regression trees can serve as predictive models. For this study, the authors restricted the input set of parameters and also branching of the tree to obtain simple trees. The trees were used mainly as a descriptive tool.

The authors also attempted to classify some process variables using fuzzy sets. Using fuzzy modeling, values of each parameter are classified into sets, typically high, medium and low. If-then-else rules are then applied to quantify (model) the relationships, for example, if Outside_Air_Temp = high, then R30 = high. These qualitative rules help understand the principal dependencies within the dataset.

Predictive Models

Predictive models can help the manufacturer react immediately when the target variable is drifting or likely to drift outside of the desired range, so that the manufacturing process can be stopped or adjusted in such a situation. A typical example of predictive modeling is the use of complex multivariate models, of which neural networks (NN) [21] and support vector machines (SVM) [22] are well-known examples. The main disadvantage of these methods is that they provide only minimum insight into the underlying relationships. One risk of developing a model using a relatively small number of data points (one season, 211 production runs) and a large number of parameters is overfitting. Additionally, prediction of target variable values outside of the rather narrow range of process variation achieved in one season is not acceptable.

A single production lot does not have to be considered as an independent experiment. Instead, the development of the target variable over lots can be understood as a time series. Forecasting in time series is a common exercise and different approaches have been investigated during the years [23]. The main attention was devoted to linear models for which the theory is known and many algorithms for model building are available. The most used linear regression methods have been the autoregressive (AR) and autoregressive moving average (ARMA) models [24]. An example of a more complex regression method is the multivariate adaptive regression splines [25]. In addition, a large number of nonlinear time series models are available. The stochastic approach to nonlinear time series that can fit nonlinear models to time series data is described in [26].

The predictive models presented in the following text are linear autoregressive models that combine both the CPP and the previous values of the target variable.

Evaluation of the Models

When evaluating descriptive models, both their lucidity and fidelity are considered. Because the lucidity is a highly

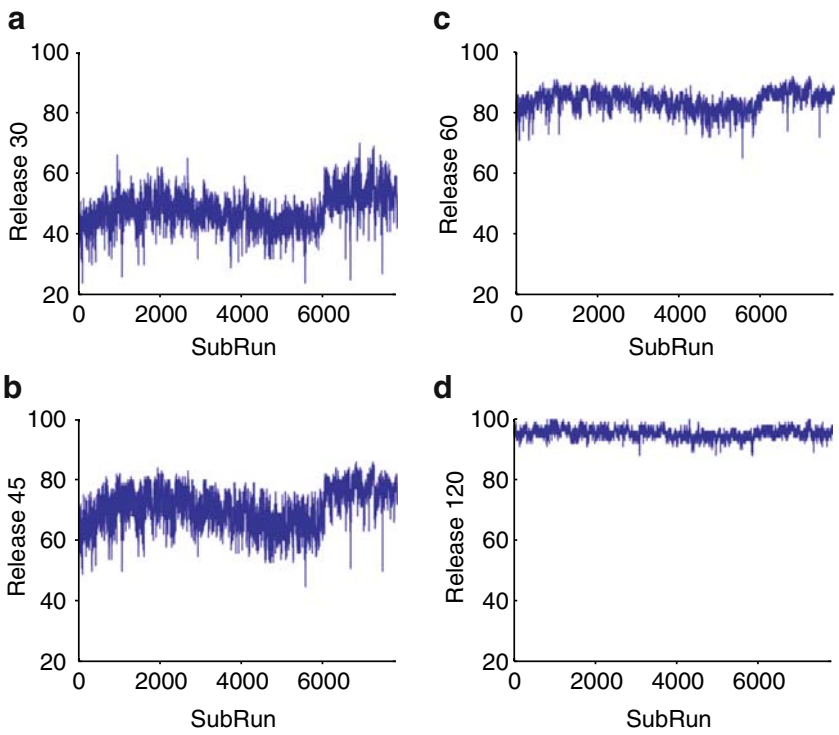

Fig. 1 Raw data for dissolution rates **a** R30, **b** R45, **c** R60 and **d** R120 as measured. At least six tablets were tested per coating SubRun; more measurements were available for re-tested SubRuns. The total number of data points is approximately 8,000

subjective criterion, the preferred structure of the descriptive models was discussed with the project team. The predictive models concern mainly the objective performance criteria such as mean absolute error (MAE) or relative standard deviation (RSD) [21].

All the predictive models have been trained and evaluated on independent training and testing sets of examples that avoid an optimistic bias in estimation of their performance.

Results and Discussion

Data Summary

Figure 1 reports the individual test results (percent released) for dissolution at each of the tested time points. The data includes all 8,000 data points that were tested. The two notable features are a large curve from data point 1 to ~5,500, and an increase in release rate from data point 6,000 onward. The exact reason for these two fluctuations was not known at the time of the analysis. The dissolution rates at 30 and 45 min were analyzed with respect to all raw material, environmental and process parameters, to determine that the notable phenomena in Fig. 1 were due to variation in ambient conditions and raw material properties.

Figure 2 is representative of the 30 min dissolution data, exhibiting average values for each coating SubRun, standard deviation and relative standard deviation. The observed variability in dissolution rate is characteristic.

CPP Ranking

This section gives an overview of CPP for Release target variables. The top parameters selected by CFS and RReliefF are shown in Table 3. The feature selection has proven that various types of parameters influence Release. Both methods more or less agree in their top identified CPP: they identify the tablet hardness as the most important parameter. All the ambient weather conditions significantly influence Release but because they are also heavily mutually dependent, only one representative is proposed for this set of parameters by both feature selection methods. CFS picks the air temperature, whereas RReliefF prefers the air pressure.

Ambient air temperature (Outside_Air_Temp) was further used as the representative parameter in empirical modeling (Model 1). By comparison, ambient air dew point (Outside_Air_Dew_Point) was also used as an alternate representative parameter to build a similar empirical model (Model 2). See Table 4 for model accuracy evaluation and Figs. 3, 4, 5, 6, 7, 8 for parameters used for modeling and for graphical illustrations of empirical modeling results.

Model Building and Dissolution (Release) Rate Modeling Summary

This study shows the impact of the most significant parameters on dissolution rates. The effectiveness of various models for predicting release rate has been evaluated.

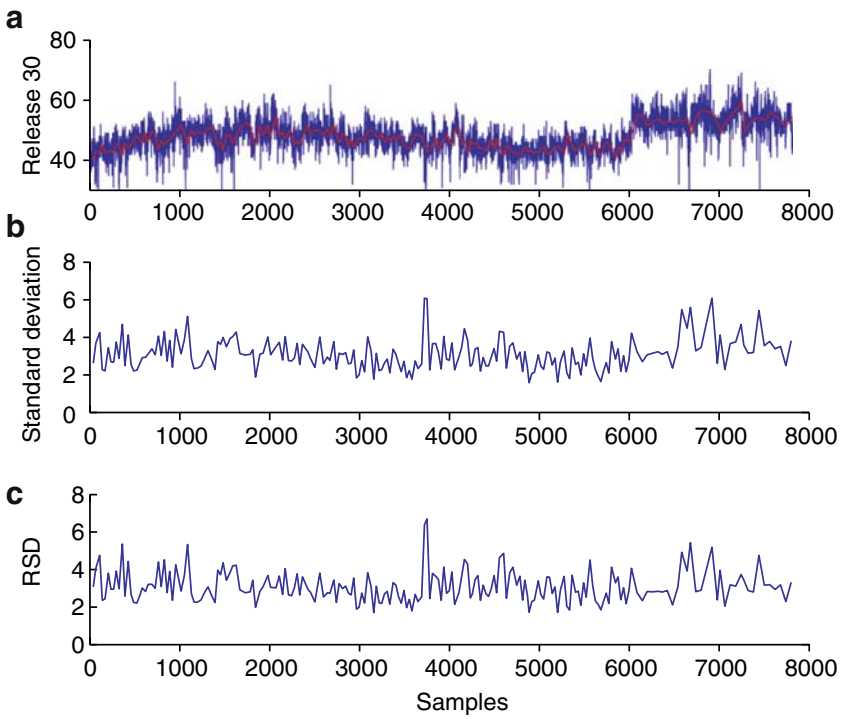

Fig. 2 **a** Raw data as measured with average R30 per coating SubRun (*red*), **b** standard deviation and **c** RSD

Table 3 List of CPP for release determined by two different sets of methods

CFS methods	RReliefF
Hardness	Hardness
Outside_Air_Temp	Outside_Air_Pressure
SRA_X	TDiff_disp_release
Precipitation	AvgOfGPower_max
IngrBulkVolume	SRA_X

Sufficient variation was found in the data for R30, R45 and R60 (Table 1) to attempt modeling these variables using other data. Empirical models, regression tree models and predictive models were used to demonstrate that these variables can be modeled with impressive accuracy. The predictive models in this article use past actual measurements of the modeled variables, other relevant current measurements and known parameter relationships to predict the future values of the modeled target variables. These models typically yielded the best overall accuracy in our experiments because they use the maximum available information when compared with other models discussed. The dissolution rate (= 'Release' or 'R') was modeled using the strongest parameters identified during CPP ranking.

Models 1 and 2, shown below, are empirical linear models with time shifted using three independent parameters to model dissolution rates after each time point.

Model 3 uses regression tree modeling techniques and Model 4 is a predictive model. These techniques were investigated for the modeling of release rates after 30, 45 and 60 min, respectively.

The list of independent parameters for all models consisted of the following:

Model 1 (empirical): Outside_Air_Temp, SRA_X, Water_Addition
Model 2 (empirical): Outside_Air_Dew_Point, SRA_X, Water_Addition
Model 3 (regression tree): Outside_Air_Temp, SRA_X
Model 4 (predictive): Outside_Air_Temp, SRA_X

Empirical Models

Figure 3 shows the results of empirical model building for Model 1; similar results were achieved using Model 2. The moving average of the dissolution rate was calculated, and the filtered results (dissolution moving average) were then estimated from all raw material, environmental and process parameters. The final model consisted of only three parameters—Outside_Air_Temp, SRA_X, Water_Addition.

Table 4 Summary of modeling results: comparison of quality of the individual models

	Description	Correlation coefficient		Results within one RSD		Results within two RSD	Model relative error	
		Filtered	ProdRun average	Filtered (%)	Actual (%)	Actual (%)	Filtered (%)	Actual (%)
Model R30								
Moving average	Moving average of ProdRun average	1.000	0.787	100.0	83.9	99.0	0	3.9
Model 1	Empirical—temperature	0.943	0.749	100.0	79.6	97.6	1.9	4.1
Model 2	Empirical—dew point	0.933	0.748	100.0	80.6	98.1	2.1	4.2
Model 3a	Regression tree—zero order Filtered	0.917	0.688	98.6	81.0	95.3	1.9	4.4
Model 3b	Regression tree—zero-order raw data Data	0.884	0.781	91.5	83.4	98.1	3.1	4.0
Model 3c	Regression tree—first-order raw data Data	0.789	0.670	80.0	72.5	97.1	4.0	4.9
Model 4a	Predictive—filtered	0.991	0.732	100.0	80.0	98.6	0.7	4.3
Model 4b	Predictive—raw data	0.943	0.757	97.6	81.0	97.6	1.2	4.2
Model R45								
Moving average	Moving average of ProdRun average	1.000	0.779	100.0	77.3	98.6	0	3.2
Model 1	Empirical—temperature	0.930	0.735	98.6	73.9	96.7	1.8	3.7
Model 2	Empirical—dew point	0.932	0.745	99.5	76.3	97.6	1.5	3.5
Model R60								
Moving average	Moving average of ProdRun average	1.000	0.743	100.0	67.3	96.7	0	1.5
Model 1	Empirical—temperature	0.940	0.694	98.6	67.3	91.9	0.7	1.7
Model 2	Empirical—dew point	0.917	0.668	92.9	64.9	91.5	1.1	2.1

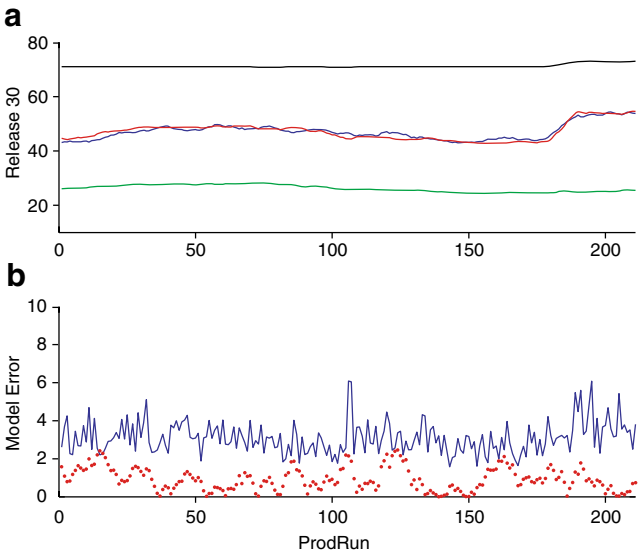

Fig. 3 **a** Model 1: R30 modeling results per ProdRun. Empirical model (*red*), filtered R30 data (*blue*), polymer property (*black*) and ambient temperature (*green*). **b** Model 1: R30 measurement (*blue*) and modeling error (*red dots*)

The accuracy of the calculated empirical model versus the known filtered results as well as the impact of one SRA_X material property and environmental conditions on the dissolution rate can be seen.

Figure 4 shows the change in temperature and dew point per production lot during the reporting period. The data suggest that a seasonal variation is occurring that appears to coincide with the observed dissolution variation (compare with Fig. 3). Considering the propensity for this material to absorb moisture readily, fluctuations in moisture content are not surprising, with the state of hydration of the polymer that provides the sustain-releasing action varying accordingly. From these data, it can be seen that temperature and dew point fluctuate together in unison, and either parameter can be used for modeling.

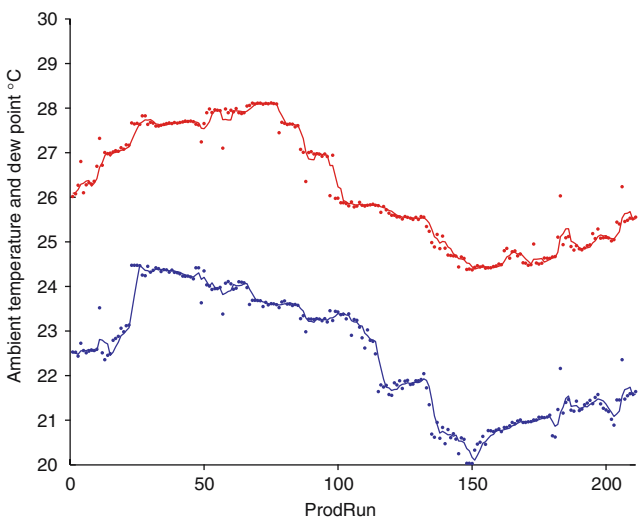

Fig. 4 Ambient air temperature (*red*) and dew point (*blue*), averaged per ProdRun

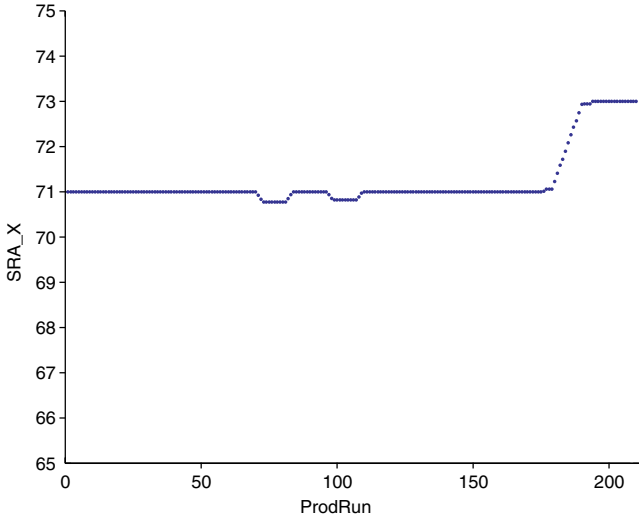

Fig. 5 Sustained-release polymer property averaged per ProdRun

Figure 5 shows the change in the polymer material property throughout the reporting period. The major shift in this property significantly impacted the dissolution rate of the product. The data can now be used to develop better, more representative raw material specifications to ensure a product with a more consistent dissolution rate.

Figure 6 shows the amount of extra granulating water that was needed above the standard quantity for certain lots. Although less obvious in its impact, this is the third most significant property affecting the release rate of the drug.

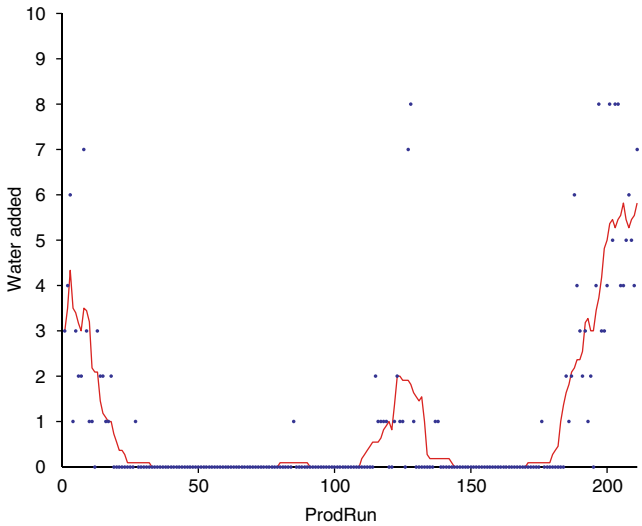

Fig. 6 Granulation water addition: total (*blue*), moving average (*red*). This chart represents total amount of water added (liters) in addition to the nominal amount of water per ProdRun

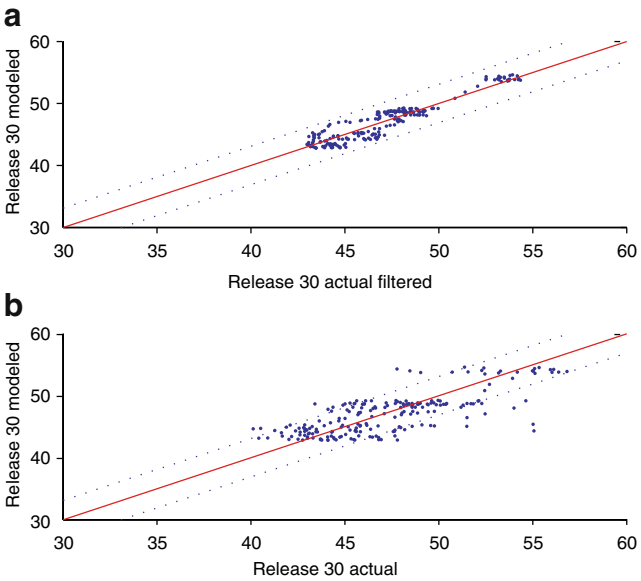

Fig. 7 **a** Model 1: actual filtered versus R30 modeled. Chart shows that all modeled results are within one standard deviation (*dotted line*) of the R30 raw data. **b** Model 1: R30 actual versus R30 modeled. Chart shows that 80% of modeled results are within one standard deviation (*dotted line*) of R30 raw data

Figures 7 and 8 show the accuracy of the models versus actual filtered results for dissolution at 30 and 45 min, respectively. For both time points, 74% or more of the modeled results are within one standard deviation of the actual dissolution rate for the raw data. The average relative standard deviation (RSD) is 6.5% for the 30 min data and 5.0% for the 45 min data. See Tables 4 and 5 for modeling results.

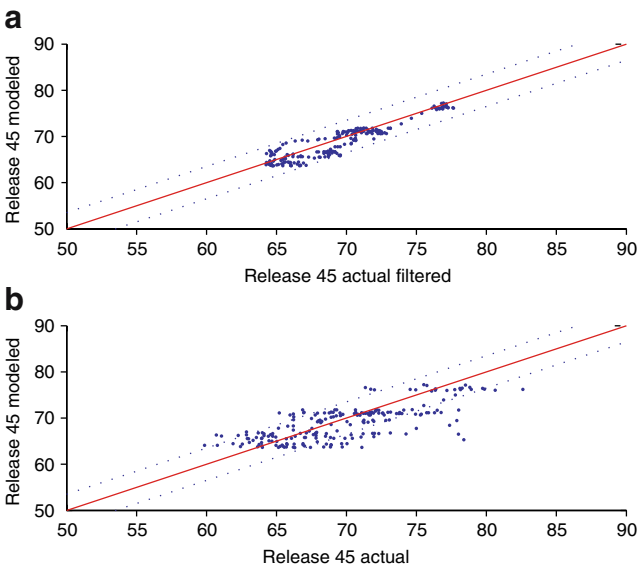

Fig. 8 **a** Model 1: R45 actual filtered versus R45 modeled. Chart shows that 99% of modeled results are within one standard deviation (*dotted line*) of R45 raw data. **b** Model 1: R45 actual versus R45 modeled. Chart shows that 74% of modeled results are within one standard deviation of R45 raw data

Table 5 Summary of modeling results: model gains for empirical models 1 and 2

	R30 gain	R45 gain	R60 gain
Model 1			
Temperature gain (%/°C)	1.8	2.3	1.6
SRA_X gain (%/°C)	5.9	6.5	3.4
Model 2			
Dew point gain (%/°C)	1.8	2.1	1.7
SRA_X gain (%/°C)	5.8	6.5	3.2

Regression Tree Models

Figure 9 displays the regression modeling results for the 30 min dissolution time point. It can be seen that the results are more variable than the previous models. Although this

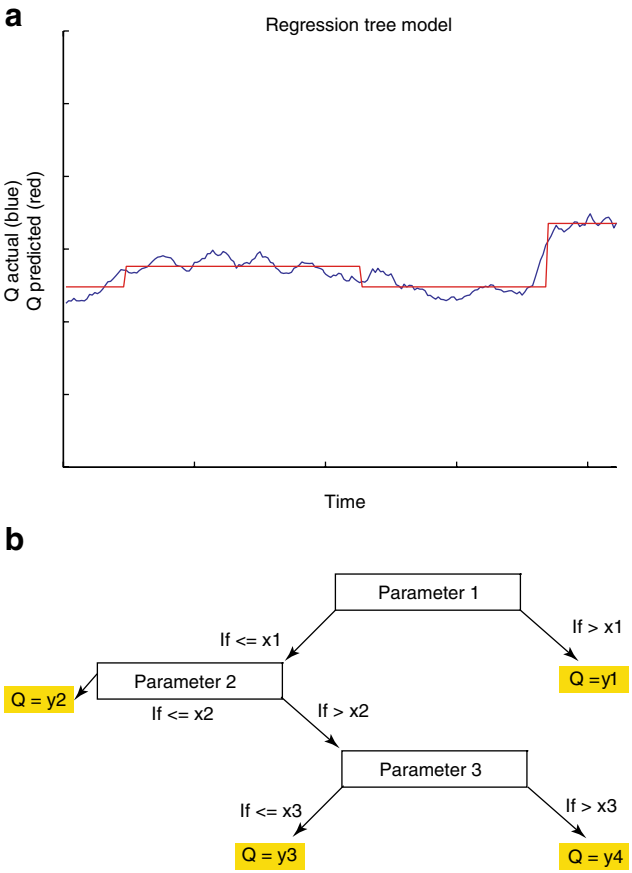

Fig. 9 **a** Model 3a: regression tree modeling results. Actual R30 results are shown in *blue* and model predicted results are shown in *red*. **b** Model 3a: regression tree model

method is not as accurate (Table 4), it should be appreciated that even such a simple model can provide valuable information about the process being modeled.

Predictive Models

Predictive models were the most accurate in predicting the dissolution rate (Table 4). Models using filtered data (Fig. 10) and raw data (Fig. 11) provided good results. The models using filtered data were more accurate (0.99 correlation coefficient, 80% of all model predictions within one measurement RSD, 98.6% within two RSD) than models using raw data (0.94 correlation coefficient, 81% of all model predictions within one measurement RSD, 97.6% within two RSD).

Correlation to Operations in Other Plants

Confirmation of the adequacy of the models was performed, whereby statistical design of experiment (DOE) methods were applied to the development and optimization of the granulation and compressing processes for this product at a second manufacturing site. Models were built from these experiments and the models terms and coefficients applied to the data from the current, approved manufacturing facility. By including the exact additional terms of the models developed from the optimization work performed at the second manufacturing site, it was possible to improve the predictability of dissolution in the currently approved manufacturing site by ~50% [27].

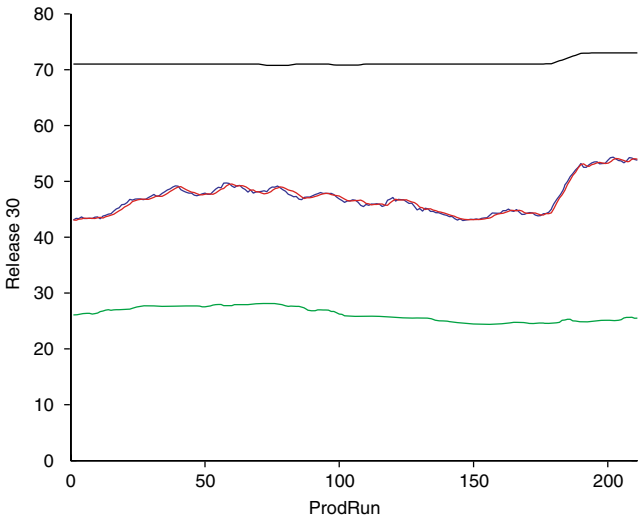

Fig. 10 Predictive modeling results using filtered dissolution results at 30 min. 100% of modeled results (*red line*) are within one standard deviation of R30 filtered data (*blue line*). Polymer raw material property fluctuation is shown in *black*. Temperature changes are shown in *green*. Correlation coefficient between modeled and actual data is high (0.991)

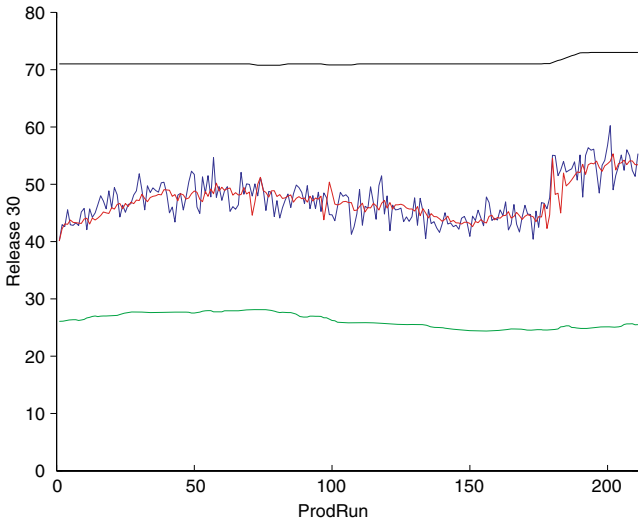

Fig. 11 Model 4b: predictive modeling results using raw dissolution data at 30 min. Polymer raw material property fluctuation is shown in *black*. Temperature changes are shown in *green*. Correlation coefficient between modeled and actual data is 0.943

Conclusions

Long-term variations of dissolution rates are caused by ambient conditions and raw material properties. Mathematical models of dissolution rates were created. The models show significant sensitivity of dissolution rates of this drug product to ambient external conditions and raw material properties.

Several automatic parameter selection and ranking methods were tested to process the data. These methods were used for automatic ranking of parameters based on their respective strength of impact on dissolution rates. The feature selection methods yielded results consistent with the current level of process understanding based on engineering principles. One particularly important property of the raw material was identified.

Understanding the impact of individual process parameter variation on critical product properties such as dissolution rate will lead to further initiatives focused on:

- Process modeling and prediction of key product properties;
- Real-time control of key product properties;
- Process optimization.

This study provides evidence of the value of process analytical technology (PAT) initiatives focused on the analysis of historical process data through the quantification of the impact of individual process and raw material parameters on key product quality attributes.

Acknowledgements The authors would like to thank Mr. Adalberto Perez and Mr. Jose Garcia from the Abbott Laboratories manufacturing facility for their guidance, hard work and dedication.

References

1. Horter D, Dressman JB. Influence of physicochemical properties on dissolution of drugs in the gastrointestinal tract. Adv Drug Deliv 2001;46:75–87.
2. D'Arcy DM et al. Evaluation of hydrodynamics in the basket dissolution apparatus using computational fluid dynamics—dissolution rate implications. Eur J Pharm Sci 2006;27:259–67.
3. Muhrer G et al. Use of compressed gas precipitation to enhance the dissolution behavior of a poorly water-soluble drug: generation of drug microparticles and drug-polymer solid dispersions. Int J Pharm 2006;308:69–83.
4. Friedrich H et al. Dissolution rate improvement of poorly water-soluble drugs obtained by adsorbing solutions of drugs in hydrohilic solvents onto high surface area carriers. Eur J Pharm Biopharm 2006;62:171–7.
5. Nerurkar J et al. Controlled-release matrix tablets of ibuprofen using cellulose ethers and carrageenans; effect of formulation factors on dissolution rates. Eur J Pharm Biopharm 2005;61:56–68.
6. von Orelli J, Leuenberger H. Search for technological reasons to develop a capsule or a tablet formulation with respect to wettability and dissolution. Int J Pharm 2004;287:135–45.
7. Hu K et al. Nanoparticle engineering processes for enhancing the dissolution rates of poorly water soluble drugs. Drug Dev Ind Pharm 2004;30:233–45.
8. Perrut M et al. Enhancement of dissolution rate of poorly soluble active ingredients by supercritical fluid processes part II: preparation of composite particles. Int J Pharm 2004;288:11–6.
9. Khan MZI. Dissolution testing for sustaining or controlled release oral dosage forms and correlation with *in vivo* data: challenges and opportunities. Int J Pharm 1996;140:131–43.
10. Lansky P. A stochastic differential equation model for drug dissolution and its parameters. J Control Release 100:267–74.
11. Li S et al. Investigation of solubility and dissolution of a free base and two different salt forms as a function of pH. Pharm Res 2004;22:628–35.
12. Fukunaka T et al. Dissolution characteristics of cylindrical particles and tablets. Int J Pharm 2006;310:146–53.
13. Witten IH, Frank E. Data mining: practical machine learning tools and techniques. 2nd ed. California: Morgan Kaufmann; 2006.
14. Huan L. Feature selection for knowledge discovery and data mining. Dordrecht: Kluwer; 1998.
15. Robnik-Sikonja M, Kononenko I. An adaptation of relief for attribute estimation in regression. In: Fisher D, editor. Machine learning: proceedings of the fourteenth international conference (ICML'97). California: Morgan Kaufmann; 1997. p. 296–304.
16. Kononenko I. Estimating attributes: analysis and extensions of RELIEF. In: Proceedings of the 1994 European conference on machine learning. 1994. p. 171–82.
17. Zaffalon M, Hutter M. Robust feature selection by mutual information distributions. Proceedings of the 14th international conference on uncertainty in artificial intelligence. 2002; 2002 cs/AI206006.
18. Hand D et al. Principles of Data Mining. Cambridge, MA: MIT Press; 2001.
19. Breiman L et al. Classification and regression trees. California, USA: Wadsworth and Brooks.
20. Quinlan JR. Learning with continuous classes. In: Adams A, Sterling L, editors. Proc. AI'92, 5th Australian joint conference on artificial intelligence. Singapore: World Scientific; 1992. p. 343–8.
21. Hastie T et al. The elements of statistical learning—data mining, inference and prediction. Berlin: Springer; 2001.
22. Vapnik V. Statistical learning theory. New York: Wiley; 1998.
23. Weigend A, Gershenfeld N, editors. Time series prediction—forecasting the future and understanding the past. Reading, MA: Addison–Wesley; 1993.
24. Box GEP et al. Time series analysis, forecasting and control. 3rd ed. Englewood Cliffs, NJ: Prentice Hall; 1994.
25. Friedman J. Multivariate adaptive regression splines. Ann Statist 1991;19:1–141.
26. Tong H. Non-linear time series. Oxford: Oxford University Press; 1990.
27. Castells X et al. Application of quality by design (qbd) knowledge from site transfers to commercial operations already in progress. J Process Anal Technol 2006;3(1):8–12.