Creation of Attribute Vectors from Spectra and Time-series Data for Prediction Model Development

Žagar, Janja and Mihelič, Jurij

Abstract: Product manufacture results in high and complex amounts of data. Most prominent and information rich are spectral and process time-series data. These data are available for every process step and raw material used for product manufacture. Due to their complexity, the use is typically very limited to research and problem solving. By applying dimensionality reduction to spectral data and specific algorithms to time-series data, attribute vectors could be created. Newly created attributes, carrying process and product details can then be used for product quality prediction. Such prediction models could replace currently used time-consuming practices.

Index Terms: ***Manufacture, prediction models, spectral data, time-series***

1. Introduction

PHARMACEUTICAL or any other industry where a product is manufactured inevitably generates enormous quantities of data. This review paper has focus on the data collected in pharmaceutical industry manufacture, but the outcomes may be applied to any other production where similar data are being collected.

The data in the manufacture environment are collected from many different sources [1]. A rough differentiation according to data source could be made as follows:

- Data from incoming raw materials;
- Data from manufacturing process;
- Data from final product analysis.

In industry, every aspect of product manufacture is highly regulated. Data are collected for every raw material batch and every process step [2]. Standard production steps include control of raw materials, combining raw materials based on a product formulation, manufacturing process steps and final analysis of the finished product.

Manuscript received June, 2019.

Žagar, J. is with the Novartis, Slovenia (e-mail: janja.zagar@novartis.com).

Mihelič, J. is with the Faculty of Computer and Information Science, University of Ljubljana, Slovenia (e- mail: jurij.mihelic@fri.uni-lj.si).

Data from incoming raw materials and final product analysis are derived from analysing the sample with pre-defined size, representing the whole batch. Every raw material batch is analysed with Near Infrared (NIR) or Infrared (IR) spectrometer probe resulting in NIR or IR spectrum. Spectrometry (i.e. NIR and IR) is at this stage used for the purpose of material identification only.

This means that each resulting spectrum is simply compared with the reference spectrum within a spectra library and if a match is found, identity is confirmed. Spectra are saved and archived and not used thereafter for anything else. Data from the manufacturing process is being collected via numerous sensors manufacturing process is equipped with.

During the entire several-hour process a number of different parameters (such as speed, flows, temperatures, etc.) are therefore being measured every few seconds and sent to different servers. This results in time-series data output for each measured process attribute. After the manufacturing process is finished, an end-product gets analysed in laboratories where its quality needs to be verified and confirmed.

Keeping in mind the above introduced high-level overview of a manufacturing process it is safe to conclude that end product quality, will depend on its building blocks and the process: Incoming raw materials and process properties.

The data in the industry presents wealth of information that could be used for gaining more knowledge about the product or the process or to predict the best manufacture process trajectory or end-product characteristics [3]. Neither of these areas have been researched to an extent to lead to applicable results. The reasons are several:

- The data collected by an industry are confidential and not readily available to computer or data scientists who could make sense of the accumulated data;
- Slow pace at which highly regulated industry such as pharmaceuticals, changes ways of working, leads to low interest by scientists;
- Raw data is in the form of Near Infrared or Infrared spectra and in the form of time-

series which presents a challenge for analysis.

1.1. Time-series Data

Time-series is a collection or sequence of numbers that represent the state of any system as a function of time or space or any other "ordering" independent variable [4]. An example consisting of several time series is presented below for main compression force parameter as a function of time for one part of the process. Process time-series data result in nonlinear dynamic. This complex, unpredictable behaviour is well known to appear in natural and physical processes such as the one presented in Figure 1.

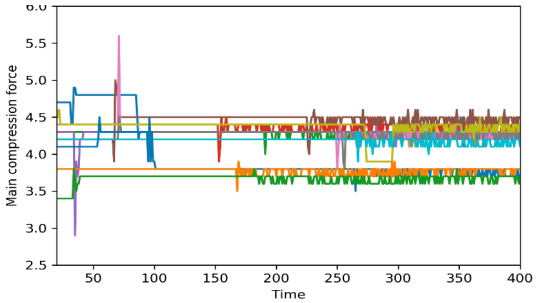

Figure 1: Example of time-series data for the selected tablet compression process parameter

1.2. Spectral Data

Spectrum is data collected by spectrometers such as near-infrared (NIR) or infrared (IR). Signal outcomes are absorbance values recorded for a whole interval of wavelength the instrument is calibrated for. Resolution of the signal collection will dictate the density of absorbance data points in the wavelength interval [5]. An example of such data output is shown in Figure 2. The approximate number of data points per one material identification spectrum is 2200 and these will describe physical and chemical attributes of measured material.

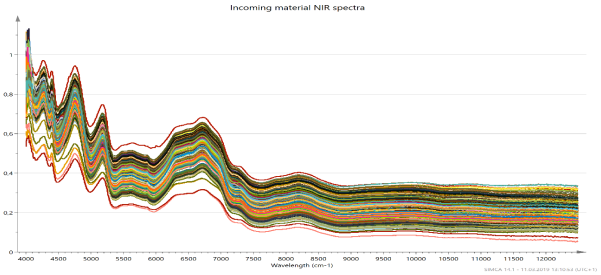

Figure 2: Near-infrared spectral data for lactose batches

1.3. Use of Data

Can we use this data in its original form? Spectra and time-series are not ideal inputs for prediction models. Is there any way to utilize these sources of data, for which we know contain valuable information? One way might be to apply embedding algorithms to spectral and time-series data and produce vector space replacing the original data.

This review will focus on, dimensionality reduction, different embedding techniques, algorithms available for representative vector formation and evaluate applicability of each for the above-described sources of data. The aim will be to show techniques available for handling such data sources and prepare new attribute vector space for prediction of end-product quality. Data used for representative analysis are derived from a randomly selected process of tablet manufacture.

2. Dimensionality Reduction

Spectra contain useful information about product chemical and physical characteristics [6]. This information is however kept in roughly 2200 variables per batch, i.e. a spectrum. A total of 2200 variables is not manageable nor is it relevant for prediction of end-product quality. Spectra are ideal candidates for so-called dimension reduction due to their homogeneous attribute responses throughout batches of an investigated material.

This homogeneous response is a consequence of material identity that dictates spectrum shape. Response intensity will differ from batch to batch only at some variables out of 2200. Furthermore, spectral attributes are also highly correlated since more wavenumbers (that is attribute/variable) describes the same material structure. This makes spectra ideal candidates for dimensionality reduction approaches, which will result in a compressed number of newly created variables.

2.1. Spectral Data Preparation

In order to reduce spectral data dimensions, a few steps need to be considered first, to prepare data for further analysis [7]. The following preparation steps should be considered before using spectra for further analysis.

Noise Removal

Parts of spectra, usually at the lowest wavenumber ranges can be highly noisy and therefore not useful for gathering information about material in question. These parts of spectra are very diverse between the batches of the same material. The highest contribution of differences between spectra for investigated material would in such case be due to noisy part of spectra. All other differences present due to chemical diversity would be insignificant in comparison. Noisy part of spectra must, therefore, be removed to enable reliable further exploratory and prediction analysis.

Pre-processing: Standard Normal Variate

Standard normal variate (SNV) pre-processing is

often used on spectra where baseline and pathlength changes cause differences between otherwise similar spectra. SNV can be used to correct for light scattering effect. This method is only applicable to spectra which have responses fairly linear in concertation. These effects that SNV removes are called multiplicative and additive.

The additive effect is constant and is observed as a baseline shift across the entire spectrum. The multiplicative effect on the other side occurs due to light scattering. The latter is in some cases regarded as noise masking other differences in spectra, in some cases, it is, however, desirable for it represents physical properties of measured substance.

After multiplicative and additive effects are removed by applying SNV pre-processing the only difference left in spectra is due to the difference in measured substance. SNV equation is applied to every measured spectrum [8]:

$$x_i = a_i + b_i z_i + e_i, \quad (1)$$

where x_i stands for a measured spectrum and z_i for the ideal one. The ideal spectrum has additive (a_i) and multiplicative inferences (b_i) which can be removed by applying SNV. The error or residual (e_i) is then mainly attributable to chemical information within the spectrum.

$$x_{icorr} = \frac{x_i - a_i}{b_i}. \quad (2)$$

A corrected spectrum is calculated as follows: from each value (parameter observation), mean of that spectrum (i.e. observation or sample) is subtracted), followed by the standard deviation for that same spectrum (i.e. row). Differences between spectra after SNV, will correspond only to differences in absorbance of material, excluding baseline shift and scattering of light.

Standardization: Mean Centering

Mean centering or center by mean, means that for every observation a corresponding variable mean is subtracted. The difference between variable mean and observation is a new point in the mean centered spectrum, i.e.,

$$x_{icentred} = x_i - \bar{x}. \quad (3)$$

Mean centering can effectively remove the mean absolute intensity from each attribute so that one can study the variability of all attributes at level zero.

Mean centering is not necessarily optimal in many cases. Mean information could be critically important in a quantitative determination of compound. It is therefore important to understand the goal and how mean centering works.

2.2. Spectral Data Analysis

Analysis which follows after spectra have been prepared involves exploratory analysis to evaluate the data and find potential groups or structures within data set we are working with. Finally, if the aim is to use this data for further prediction analysis, dimensionality reduction is required, since we cannot proceed with 2200 attributes (also called variables in this paper) per one observation (i.e. batch).

Principal Component Analysis

Principal component analysis (PCA) is a method of choice for initial exploratory analysis of large spectral data. It can also be utilized for dimension reduction. PCA is a method that aims to detect trends and clusters in large data sets by generating a series of fewer, new latent variables, called principal components (PC) [9]. These components are formed mathematically by locating and extracting the principal sources of variability within the dataset. They are formed from the original variables and observations by combining them into a simpler linear combination of those original variables.

Pre-requisite for PCA to work properly is inter-relation or presence of correlation between original variables. Original data should be normal or approaching to normal and should also exhibit some variability since PCA is designed to search for major sources of variability [10].

Incoming raw material spectral data was used to demonstrate spectral data preparation and PCA analysis. Spectral data had SNV and mean centering applied. The noise was not removed but should be attempted in future analysis and explored whether subsequent PCA analysis performs better. PCA applied to raw spectral data and pre-processed spectral data differs significantly. The latter means that baseline and measurement pathlength changes cause main differences and other subtle variation between spectra observations (batches) did not come through in PCA.

After pre-processing is applied differences between batches are apparent, see Figures 3, 4, 5, and 6. There are some structures observed and grouping of same colour (i.e. coloured according to batches). New observations are now called scores and new variables PCs. Ideally first few (e.g. two) PCs will explain for most variability in the dataset and could be used as new variables for the dataset, replacing 2200 variables we started with.

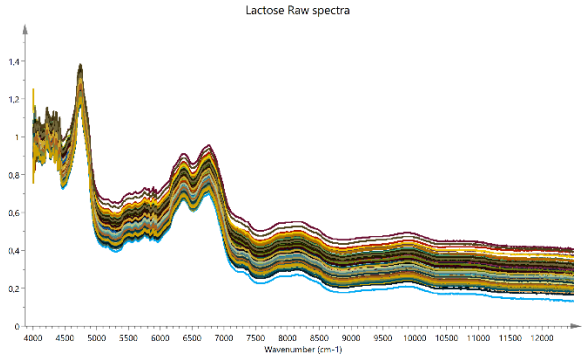

Figure 3: Raw spectra before application of pre-processing

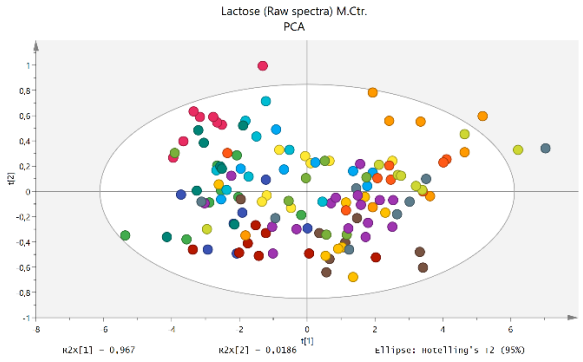

Figure 4: Principle component analysis using spectra presented in Figure 3

Euclidean Distance

Euclidian distance is a good choice for pre-processed spectral data that correspond to the same material chemistry-wise and only differ slightly in physical properties or content of certain compounds.

In the case presented in this paper, an optimal raw material batch would be defined that led to the highest quality product at the end. Spectrum ($x_{opt.}$) corresponding to this particular batch would be chosen for Euclidean distance calculation for all other spectra. The result would be one new variable replacing initial 2200. Distance calculation between observation spectrum (x) and optimal spectrum (z) is presented below [11].

$$D_i = \sqrt{(x_1 - z_1)^2 + (x_2 - z_2)^2 + \cdots + (x_n - z_n)^2} \quad (4)$$

3. Attribute Vector Creation

Process time-series do not have predicted dynamics and batches do not start the process at precisely the same point as was true for spectral data. Instead, different lag times are usually present and unknown dynamics. Time-series also need to have their initial attribute number reduced from as high as 10000 per one time-series parameter per batch to as few as possible to cover all relevant variability and process properties.

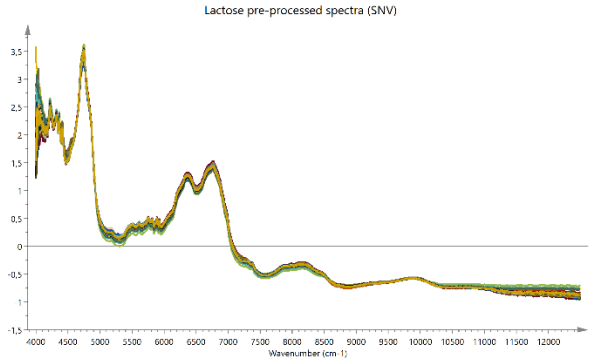

Figure 5: Pre-processed spectra (SNV and Mean centering)

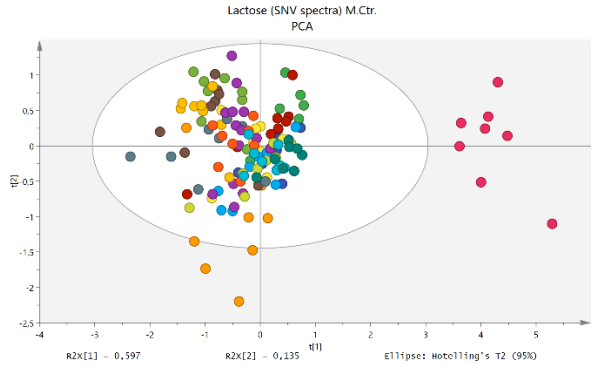

Figure 6: Principle component analysis with pre-processed spectra from Figure 5

The approach would again include data preparation, followed by new attribute vector preparation.

3.1 Time-series Data Preparation

Process data have different starting points and durations of process stages, as well as different potential interruptions as seen in Figure 7. The data needs to be prepared to enable meaningful analysis.

In case of tablet compression process, we can work with process time-series that indicate the starting point and ending point of the process. This particular time-series would then be used to adjust the starting point of all other time-series for that particular batch.

Such an example is a tablet press speed time-series parameter, presented below for randomly selected batches. This time-series only has value higher than 0 when the process is running and 0 when the process is stopped or interrupted. It is therefore ideal to define the actual process starting point across all time-series for each batch.

Besides defining the starting and ending point of the process, larger time interruptions would also need to be removed/cleaned from time-series. The latter needs to be evaluated by a domain expert, who can determine which interruptions are part of standard operation (e.g. operator's shifts, weekends) and need to be removed. Some interruptions, on the other hand, indicate process issues and should be monitored

and included as weighed attribute in new vector space.

The domain expert is needed to:

- Evaluate the removal of entire outlier batches due to related issues that led to batch rejection.
- For each time-series type per batch (i.e. dynamic process parameter), specifics that have an impact on product quality need to be defined and methods for capturing those specifics determined.
- Evaluate which time-series should be compared between observations/batches and what should these be compared with: time-series of a specific batch or an average time-series across all batches or a different comparison approach.

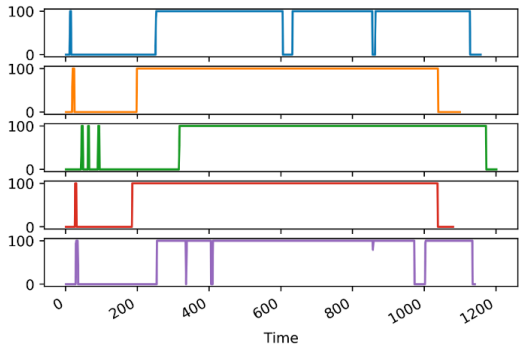

Figure 7: Tablet press speed time-series data for five randomly selected batches/observations

3.2. Creation of New Attributes

Once time-series have been prepared for all batches (observations) included in analysis, new attributes need to be created. Each time-series consists of approximately 10000 initial attributes which can't be worked with (similarly as with spectra). Simple dimensionality reduction with method like PCA will not be sufficient in this case due to complexity of time-series outputs. New attribute vectors need to be tailored separately for each parameter based on expert knowledge and application of optimal algorithms.

Dynamic Time Warping (DTW)

A simple transformation of time-series datasets into distances may be succeeded by using so-called Dynamic Time Warping (DTW) [12]. The method is used mainly for finding differences between sets of time-series.

As opposed to Euclidian distance which defines the distance between each pair (i.e. one on one) the DTW instead looks for best alignment between points in the two time series. DTW, therefore, can compare two to three points with one in comparative time series to result in best alignment between the two. This method allows to better compare time series that are not exactly the same but are similar [13]. See also Figure 8.

Distances are calculated via distance matrix where values of two plots are used to calculate distances between points as shown in Equation (5) [12]. Each time series gets new value relative to the selected time series it is compared with.

1. Initial condition: $g(1,1) = d(1,1)$
2. Repeat: for $1 \leq i \leq n, 1 \leq j \leq m$

$$g(i,j) = min\begin{cases} g(i-1,j) + d(i,j) \\ g(i-1,j-1) + d(i,j) \\ g(i,j-1) + d(i,j) \end{cases}$$

3. Finish: $D(X_i,X_j) = g(n,m)/(n+m)$ (5)

where $d(i,j)$ stands for the distance between two cells or observations between the two time series with length n and m and $g(i,j)$ is a newly calculated distance including warping factor, i.e. minimum distance considering neighbouring observations. $D(X_i,X_j)$ is finally calculated and normalized distance between two time series.

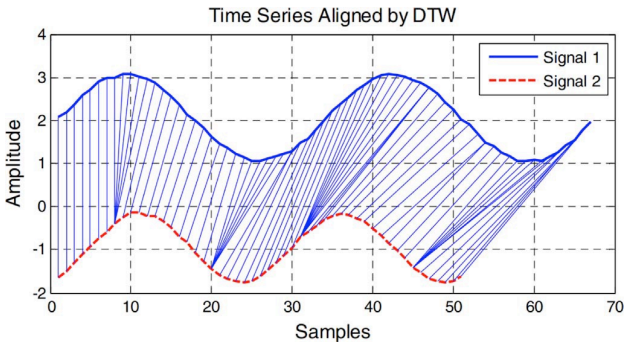

Figure 8: DTW method for two time-series [13]

3.3 Application of Simple Algorithms

Simple algorithms should be applied to extract within time-series variability and properties. Proposal of time-series features that should be extracted and included in new attribute vector space are listed below. They were tailored based on expert knowledge of process and product as well as recommended time-series algorithms [14]. All of these can indicate on process successfulness and by that product quality outcome for the investigated batch. We give a few examples below:

- Number of equipment stoppages during normal process run, weighed by the duration of interruption.
- Number of attempted equipment setups before the process started, weighed by their duration indicates either on ease of material handling or operator's experience. Both can have an impact on process successfulness.
- Rate of increase/decrease of main physical parameters (time-series) compared to average value throughout the entire batch.
- Number of events when main physical parameters (i.e. time-series) increase by 15% from the set target value.

- Number of events above fixed upper tolerance value for selected physical parameters.
- Average, median, min and max values for selected physical parameters.
- DTW comparison between observed time-series and an average across all observations included in the analysis; for all time-series types.
- The number of changes from the initial set target value, weighed by the extent of change for certain physical parameters (mainly for speed time-series).

Number of attributes per one time-series could be as large as 10 or more based on the above-approximated summary. The number of time-series per batch is at least 8 or more. The latter is a considerable reduction from approximately 8x 10000 attributes per batch only contributed by time-series. Approximately 80 attributes could however still be quite a large number for some predictive modelling methods.

In cases where researcher evaluates new attribute vector space is too large for selected prediction methods, newly defined attribute space can be further reduced by testing for significance. The use of the null hypothesis (H0) is one way to test it. H0 stating there is no impact of the selected (newly created) attribute on quality of the product we are attempting to predict. If H0 is accepted, the attribute is rejected [14]. The loop is run for each newly created attribute and only most significant ones are then used for the final step: prediction of product quality.

4. Predictive Modelling

The ultimate aim is to use newly created attributes from time-series and spectral data as well as other relevant simple attributes related to observations (e.g. simple one-point analysis of materials), for prediction of end-product quality. Prediction models will depend on the final number of attributes we need to include.

Since end-product quality is a known attribute, available for all observations included in the analysis, supervised learning methods are applicable. Some of the prediction methods that can be considered are mentioned below. A detailed description is not included for these are well known and established approaches researchers are familiar with.

Linear Regression

Parametric test with a relatively simple connection between parameters and target variable we want to predict. Here, a linear response between attributes and target is assumed, which may cause an issue if the linear response is non-existent. Special care must be taken to avoid over-fitting by applying optimal regularization level [15]. An example of the linear regression model is presented below, considering only simple attributes (see Figure 9). Further optimization and addition of special attributes from spectral and time-series data would improve the accuracy of the model further.

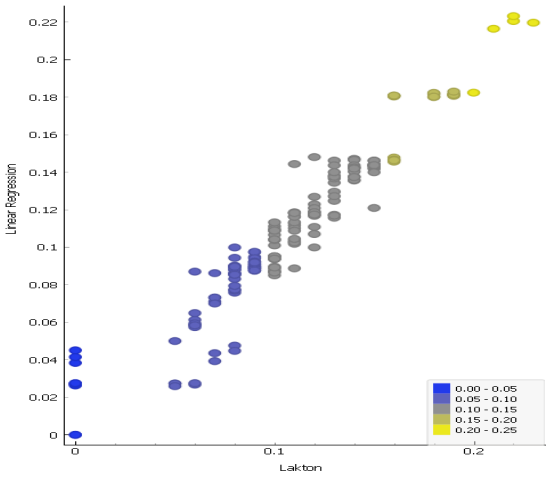

Figure 9: Linear regression model for "Lakton" prediction

K-Nearest Neighbors

If linear dependence between attributes and dependent variable (target) is not present and none of the applied transformations do not improve linear regression outcome, a non-parametric method will probably be needed. An example of the non-parametric test is the k-nearest neighbor (kNN) which uses an average of k-neighbors as a prediction value. Neighbors are defined by calculating distances (e.g. Euclidean distance) between all observations.

Artificial Neural Networks

When dataset consists of a very large number of observations (i.e. thousand and more) a good choice might be artificial neural networks (ANN) which combine linear and non-linear functions by adding hidden layers between input and output layers. ANNs are especially good in self-optimization since prediction errors are used as weights for functions in incoming layers [16].

Prediction methods will however only perform well if input data is optimal and represents a dependent variable (e.g. end-product quality) well. A researcher should, therefore, consider optimizing the generation of attributes itself. Attributes creation as described in this paper should be linked with the prediction model and a loop created to search for such algorithm parameters to yield optimal prediction outcome.

5. Conclusion

Time-series and spectral data are heavily represented in manufacture industries. These

data are usually collected and never used for any particular reason beyond the initial identification of materials (spectral data) and investigation of exceptional process failures (time-series data) [17]. What if companies could utilize this data for something that will add value to their product and business? What if we could use this data to accurately predict end-product quality and avoid testing manufactured product in laboratories after completed production? Laboratory testing can add up to 300% of the time to overall product throughput time. This time is restlessly aimed to be reduced.

Spectral and time-series data have too high dimensionality for direct prediction modelling. A lot of information in this raw inputs is also not relevant for chosen end-product quality and would present noise in the dataset. By employing dimensionality reduction techniques and customized attribute vector creation, resulting in a reduced (and relevant) attribute space can be used for prediction modelling with the chosen method.

Research needs to be focused on optimal attribute vector creation. Only successful new attribute space creation from complex input data will result in the prediction model that could replace current laboratory testing for the selected product. Having end-product quality available at the time of manufacture completion would revolutionize any industry where such an approach is implemented.

REFERENCES

[1] Gams M., Horvat M., Ožek M., Luštrek M., and Gradišek A., 'Integrating Artificial and Human Intelligence into Tablet Production Process', AAPS PharmSciTech, 2014, vol. 15, no. 6, pp. 1447–1453.

[2] Lawrence X. Y., 'Pharmaceutical quality by design: product and process development, understanding, and control', Pharmaceutical research, 2008, vol. 25, no. 4, pp. 781–791.

[3] Huang J. et al., 'Quality by design case study: an integrated multivariate approach to drug product and process development', International journal of pharmaceutics, 2009, vol. 382, no. 1–2, pp. 23–32.

[4] Konar A. and Bhattacharya D., Time-series prediction and applications. Springer, 2017.

[5] Roggo Y., Chalus P., Maurer L., Lema-Martinez C., Edmond A., and Jent N., 'A review of near infrared spectroscopy and chemometrics in pharmaceutical technologies', Journal of pharmaceutical and biomedical analysis, 2007, vol. 44, no. 3, pp. 683–700, 2007.

[6] S. Laske, A. Paudel, and O. Scheibelhofer, 'A Review of PAT Strategies in Secondary Solid Oral Dosage Manufacturing of Small Molecules', J. Pharm. Sci., vol. 106, no. 3, pp. 667–712.

[7] Rinnan A., Van Den Berg F., and Engelsen S.B., 'Review of the most common pre-processing techniques for near-infrared spectra', TrAC Trends in Analytical Chemistry, 2009, vol. 28, no. 10, pp. 1201–1222.

[8] Bi Y., et al., 'A local pre-processing method for near-infrared spectra, combined with spectral segmentation and standard normal variate transformation', Analytica Chimica Acta, 2016, vol. 909, pp. 30–40.

[9] Rajalahti T. and Kvalheim O. M., 'Multivariate data analysis in pharmaceutics: a tutorial review', International journal of pharmaceutics, 2011, vol. 417, no. 1–2, pp. 280–290.

[10] Košmelj K., 'Metoda glavnih komponent: osnove in primer', Acta agriculturae Slovenica, 2007, vol. 89, no. 1, pp. 159–172.

[11] Dasgupta S., Papadimitriou C. H., and Vazirani U. V., Algorithms, McGraw-Hill Higher Education, 2008.

[12] Hayashi A., Mizuhara Y., and Suematsu N., 'Embedding time series data for classification', in Machine Learning and Data Mining in Pattern Recognition, Proceedings, Springer-Verlag Berlin, 2005, vol. 3587, pp. 356–365.

[13] Lu B., Xu S., Stuber J., and Edgar T. F., 'Constrained selective dynamic time warping of trajectories in three dimensional batch data', Chemometrics and Intelligent Laboratory Systems, 2016, vol. 159, pp. 138–150.

[14] Christ M., Kempa-Liehr A. W., and Feindt M., 'Distributed and parallel time series feature extraction for industrial big data applications', arXiv:1610.07717, 2016.

[15] Bickel P. J. et al., 'Regularization in statistics', Test, 2006, vol. 15, no. 2, pp. 271–344.

[16] Basheer I. A. and Hajmeer M., 'Artificial neural networks: fundamentals, computing, design, and application', J. Microbiol. Methods, 2000, vol. 43, no. 1, pp. 3–31.

[17] Klemenčič J., Mihelič J., 'Application of Algorithms and Machine Learning Methods in Pharmaceutical Manufacture', The IPSI BgD Transactions on Internet Research, 2019, vol. 15, pp. 19-26.

Janja Žagar received her Master of Science degree at King's College London in the field of Pharmaceutical Technology. She has gained pharmaceutical industry insight by working at GlaxoSmithKline and Novartis where she is currently employed. Her area of expertise lies in optimisation of manufacturing processes by innovative data science applications.

Jurij Mihelič received his doctoral degree in Computer Science from the University of Ljubljana in 2006. Currently, he is with the Laboratory of Algorithmics, Faculty of Computer and Information Science, University of Ljubljana, Slovenia, as an assistant professor. His research interests include algorithm engineering, combinatorial optimization, computer systems and software.