Doctoral Thesis

# Development of practical soft sensors for water content monitoring in fluidized bed granulation considering pharmaceutical lifecycle

Keita Yaginuma

Graduate School of Informatics, Kyoto University

March 2022

**List of publications contributing to the present thesis**

Keita Yaginuma, Shuichi Tanabe, Takuya Miyano, Hiroshi Nakagawa, Satoshi Suzuki, Shuichi Ando, Manabu Kano. Scale-free soft sensor for monitoring of water content in fluid bed granulation process. *Chem. Pharm. Bull.* (2020) 68, 855-863

Keita Yaginuma, Shuichi Tanabe, Hirokazu Sugiyama, Manabu Kano. Prediction performance and economic efficiency of soft sensors for in-line water content monitoring in fluidized bed granulation: PP-based model vs. NIRS-based model. *Chem. Pharm. Bull.* (2021) 69, 548-556

Keita Yaginuma, Shuichi Tanabe, Manabu Kano. Gray-box soft sensor for water content monitoring in fluidized bed granulation. *Chem. Pharm. Bull.* (2022) 70, 74-81

**Conference participation – oral presentations**

Keita Yaginuma, Takuya Miyano, Hiroshi Nakagawa, Tomoyuki Watanabe, Hidemi Minami, Manabu Kano. Exploitation of process parameters having small differences between granulators for monitoring of water content during fluid bed granulation. The Society of Chemical Engineers, Japan (SCEJ) 82$^{nd}$ Annual Meeting, Tokyo, Japan (2017)

Keita Yaginuma, Shuichi Tanabe, Manabu Kano. Scale-free monitoring of water content during fluid bed granulation without spectroscopy. 18$^{th}$ Asian Pacific Confederation of Chemical Engineering Congress, Sapporo, Japan (2019)

**Awards**

Keita Yaginuma, Takuya Miyano, Hiroshi Nakagawa, Tomoyuki Watanabe, Hidemi Minami, Manabu Kano. Exploitation of process parameters having small differences between granulators for monitoring of water content during fluid bed granulation. Technical Achievement Award of the Division of Systems, Information, and Simulation (SIS) in SCEJ 82nd Annual Meeting (2017)

Kensaku Matsunami, Shuichi Tanabe, Keita Yaginuma, Hirokazu Sugiyama. Economic evaluation of solid drug product manufacturing processes considering the choice of continuous manufacturing. Technical Achievement Award of the Division of SIS in SCEJ 84th Annual Meeting (2019)

**List of publication not contributing to the present thesis**

Kensaku Matsunami, Fabian Sternal, Keita Yaginuma, Shuichi Tanabe, Hiroshi Nakagawa, Hirokazu Sugiyama. Superstructure-based process synthesis and economic assessment under uncertainty for solid drug product manufacturing. *BMC Chem. Eng.* (2020) 2, 6

# Table of Contents


**1. Introduction ..... 1**

**1.1. Manufacturing process of pharmaceutical tablets ..... 1**

**1.2. Paradigm shift to assure higher-quality pharmaceuticals ..... 3**

**1.3. Soft sensor ..... 4**

**1.4. Pharmaceutical lifecycle ..... 7**

**1.5. Aim of work and thesis outline ..... 9**

**2. General information ..... 12**

**2.1. Formulation ..... 12**

**2.2. Granulators ..... 12**

**2.3. PPs measurement ..... 13**

**2.4. Experimental data for calibration and validation ..... 13**

**2.4.1. Formulation A ..... 14**

**2.4.2. Formulation B ..... 14**

**2.5. NIR spectra ..... 18**

**2.5.1. Measurement ..... 18**

**2.5.2. Preprocessing ..... 18**

**2.5.3. Wavenumber selection ..... 18**

**2.6. Reference measurement ..... 18**

**2.7. Regression method ..... 19**

**2.7.1. LW-PLSR ..... 19**

**2.7.2. PLSR ..... 22**

**2.8. Prediction performance evaluation ..... 22**

**2.9. Evaluation of LW-PLSR model ..... 23**

**3. Scale-free PP-based Black-box Soft Sensor ..... 25**

**3.1. Introduction ..... 25**

**3.2. Materials and Methods ..... 25**

**3.2.1. Materials ..... 25**

**3.2.2. Measurement of PPs, NIR spectra, and reference ..... 26**

**3.2.3. Calibration and validation datasets ..... 26**

**3.2.4. Input variable selection ..... 26**

**3.2.4.1. PPs ..... 26**

**3.2.4.2. NIR spectra ..... 27**

**3.2.5. Regression method ..... 27**

**3.2.6. Prediction performance evaluation ..... 28**

**3.2.7. Evaluation of LW-PLSR model ..... 28**

**3.3. Results and discussion** ..... 28
**3.3.1. PP-based black-box models** ..... 28
**3.3.1.1. Input variable selection** ..... 28
**3.3.1.2. Regression method** ..... 30
**3.3.2. NIRS-based black-box models** ..... 33
**3.3.2.1. Preprocessing** ..... 33
**3.3.2.2. Wavenumber selection** ..... 34
**3.3.2.3. Regression method** ..... 34
**3.3.3. Comparison of the prediction accuracy** ..... 35
**3.4. Summary** ..... 36
**4. Prediction Performance and Economic Efficiency of Black-box Soft Sensors: PP-based Model vs. NIRS-based Model** ..... **37**
**4.1. Introduction** ..... 37
**4.2. Materials and Methods** ..... 38
**4.2.1. Materials** ..... 38
**4.2.2. Measurement of PPs, NIR spectra, and reference** ..... 38
**4.2.3. Calibration and validation datasets** ..... 39
**4.2.4. Input variables** ..... 40
**4.2.4.1. PP-based models** ..... 40
**4.2.4.2. NIRS-based models** ..... 40
**4.2.5. Regression method** ..... 41
**4.2.6. Prediction performance evaluation** ..... 41
**4.2.7. Development cost of PAT** ..... 41
**4.3. Results and discussion** ..... 43
**4.3.1. Prediction accuracy** ..... 43
**4.3.1.1. PP-based LW-PLSR models** ..... 43
**4.3.1.2. NIRS-based LW-PLSR models** ..... 49
**4.3.2. Comparison of development cost of PAT** ..... 52
**4.4. Summary** ..... 54
**5. Gray-box Soft Sensor** ..... **55**
**5.1. Introduction** ..... 55
**5.2. Materials and Methods** ..... 55
**5.2.1. Materials** ..... 55
**5.2.2. Measurement of PPs and reference** ..... 55
**5.2.3. Calibration and validation datasets** ..... 56
**5.2.4. Heat and mass balance model (white-box model)** ..... 56

**5.2.4.1. Equations for water content calculation** ..... 57
**5.2.4.2. Fitting parameters optimization** ..... 58
**5.2.5. Gray-box models** ..... 60
**5.2.5.1. Parallel gray-box model** ..... 60
**5.2.5.2. Serial gray-box model** ..... 61
**5.2.5.3. Combined gray-box model** ..... 62
**5.2.6. Prediction performance evaluation** ..... 63
**5.2.7. Evaluation of LW-PLSR model** ..... 63
**5.3. Results and discussion** ..... 63
**5.3.1. Model building** ..... 63
**5.3.1.1. Heat and mass balance model (white-box model)** ..... 64
**5.3.1.2. Parallel gray-box model** ..... 64
**5.3.1.3. Serial gray-box model** ..... 68
**5.3.1.4. Combined gray-box model** ..... 68
**5.3.2. Applicability to real operating data on commercial scale** ..... 69
**5.4. Summary** ..... 74
**6. Conclusion** ..... **75**
**Acknowledgements** ..... **80**
**Abbreviations** ..... **81**
**References** ..... **83**

# 1. Introduction

## 1.1. Manufacturing process of pharmaceutical tablets

A drug product is composed of one or more active pharmaceutical ingredients (APIs) and several functional excipients, such as diluent, disintegrant, binder, and lubricant, and manufactured through various unit operations. A representative manufacturing flow of pharmaceutical tablets consists of a granulation process, a blending process, a tableting process, and a film coating process. The granulation process to transform powder mixture into granule improves powder's properties, such as flowability, handling, bulk density, dust formation, resistance to segregation, and solubility.[1)] The blending process to mix the granules and a lubricant in a tumble mixer is conducted to prevent sticking or die friction trouble in a tableting process.[2)] In the tableting process, tablets are manufactured by compression of the lubricated granules on a table rotating at high speed. The film coating process provides better appearance, distinguishability, and easier swallowing and protects the drug product against environmental influences such as humidity and light.[3)]

Among these unit operations, the granulation process is critical because the granule properties, such as water content, particle size distribution, and bulk density, significantly impact the manufacturability in downstream processes and eventually affect the final product qualities, such as drug release and stability. Fluidized bed granulation, one of the most applied granulation processes, is complex[4)] because it has more process parameters (PPs) than other granulation processes: for example, spray rate, inlet air volume, inlet air temperature, inlet air humidity, nozzle air pressure, and nozzle height.[5)] A schematic diagram of a top-spray fluidized bed granulator is shown in Figure 1.1. The fluidized bed granulation process is divided into two phases: the granulation phase and the drying phase.

In the granulation phase, a binder solution is sprayed onto the particles fluidized by the heated inlet air from the bottom of the granulator. After the granulation phase, wet granules are dried by the heated inlet air. The water content profile during granulation is a crucial factor affecting the granule properties, such as particle size distribution and bulk density, which in turn could impact the tablet properties, such as hardness and disintegrating time.[6] Conventional off-line water content measurement methods, such as loss on drying (LOD) and Karl Fischer, are unsuitable for real-time process monitoring. Thus, the prediction model to estimate water content has been widely utilized for in-line water content monitoring.[7]

In the pharmaceutical industry, the scaling-up of the fluidized bed granulation process from small scale to commercial scale has been conducted empirically,[8] whereas the complicacy of this process usually makes it challenging. Hence, the first-principle models, such as heat and mass balance models[9-14] and population balance models (PBMs),[15-18] have been used to enhance process understanding of fluidized bed granulation or drying.

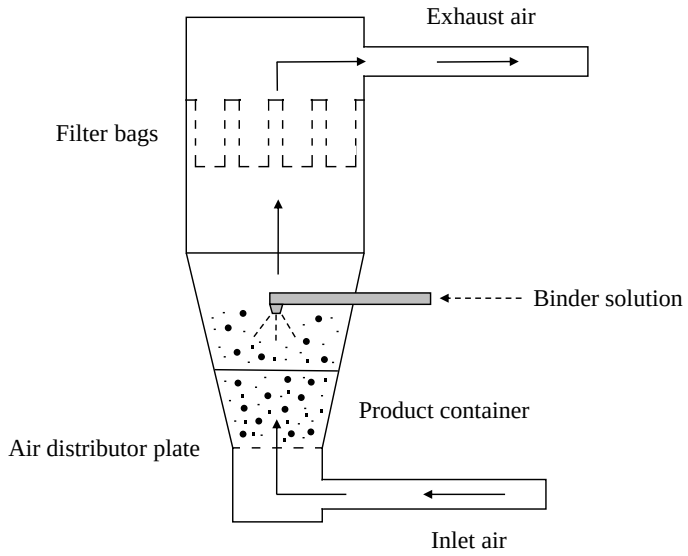


Figure 1.1. Schematic diagram of a top-spray fluidized bed granulator.

## 1.2. Paradigm shift to assure higher-quality pharmaceuticals

The PPs in each unit operation and material attributes (MAs) of intermediate products, such as granules and uncoated tablets, could affect the final product qualities. The relationship between PPs, MAs, and final product qualities is complex due to their interactions. Before commercial manufacturing, the process validation is performed to demonstrate that the developed manufacturing process can consistently deliver quality products. Traditionally, the process validation is completed by successfully manufacturing three consecutive commercial-scale lots at the target operating conditions. However, this traditional approach is insufficient to establish a robust process control strategy against the variations of PPs or raw material properties.

To assure pharmaceutical quality at a higher level, several guidelines[19-25] regarding Quality by Design (QbD)[26,27] and process analytical technology (PAT)[19] were published by the Food and Drug Administration (FDA), the European Medicines Agency (EMA), and the International Conference on Harmonisation of Technical Requirements for Registration of Pharmaceuticals for Human Use (ICH). QbD is a systematic approach based on quality risk management. The following steps establish a control strategy based on QbD. At first, we define a Quality Target Product Profile (QTPP), which is a prospective summary of the quality characteristic of a final product,[25] and Critical Quality Attributes (CQAs), which are physical and chemical properties necessary to be within an appropriate range to assure the desired product quality.[25] Then, we determine Critical Material Attributes (CMAs) and Critical Process Parameters (CPPs), which affect CQAs, through experiments conducted by changing manufacturing conditions to evaluate the influence of MAs and PPs on CQAs. At last, we set the acceptable ranges of CMAs and CPPs to satisfy the criteria of CQAs consistently, which results in the establishment

of a control strategy to assure consistency of product quality through commercial manufacturing. Hence, QbD promotes product and process understanding and contributes to selecting the target operating condition for commercial manufacturing based on a scientific rationale. PAT is a system for designing, analyzing, and controlling manufacturing processes through real-time monitoring,[25] which enhances the process control strategy based on QbD. In other words, feedback or feedforward control of CPP is performed using in-line monitoring of CMA, which contributes to the consistency of CQA. Besides, PAT is valuable to facilitate process understanding in a process development study. Therefore, PAT plays an essential role throughout the pharmaceutical lifecycle.

## 1.3. Soft sensor

The in-line monitoring methods are divided into two types: hard sensors and soft sensors. Hard sensors are electromechanical or optical sensors to measure values directly.[28] On the other hand, soft sensors are prediction models, and they can estimate specific attributes, which cannot be measured in real-time due to technological or economical reasons.[29] Thus, soft sensors have been widely applied to various industries, such as chemical,[29] paper/pulp,[29] steel,[29,30] petrochemical,[30] semiconductor,[30] and pharmaceutical.[30] In the pharmaceutical industry, soft sensors using near-infrared spectroscopy (NIRS) have been extensively applied to various processes, such as fluidized bed granulation or drying,[31-34] blending,[35-37] tableting,[38] and film coating.[39,40] NIRS has the ability to predict various chemical and physical qualities rapidly and nondestructively.[41-43] Additionally, NIRS-based monitoring methods are generally robust enough to counter the manufacturing scale change. However, NIRS has a known

drawback of the high initial investment cost of the near-infrared (NIR) spectrometer and the probe.

Soft sensors are categorized into three types, *i.e.*, white-box, black-box, and gray-box models,[44] as shown in Figure 1.2. White-box models are based on first-principle, which means that they are descriptive and easy for industry operators to interpret.[45] Several studies on white-box models of fluidized bed granulation or drying have been reported.[9-14] However, white-box models have a known drawback; sometimes they cannot grasp the actual dynamics of a complex industrial process, which causes deterioration of their prediction accuracy.[45] For example, Tanino *et al.*[10] demonstrated that the prediction accuracy of the heat and mass balance model could deteriorate depending on the formulations. Wang *et al.*[13] reported that the mass balance model predicted higher values of granule water content than the reference values.

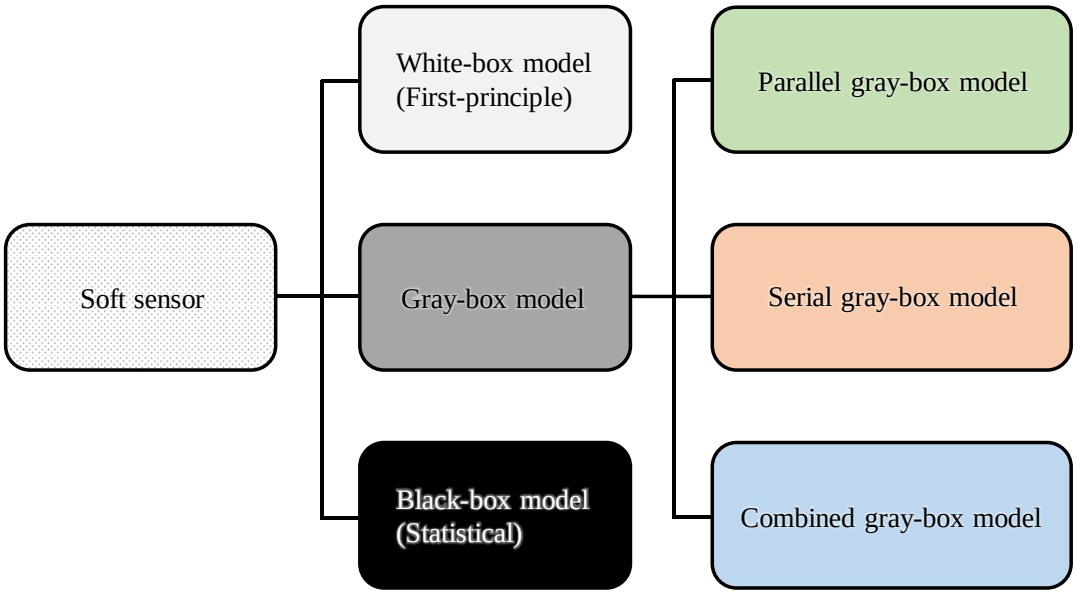


Figure 1.2. Classification of soft sensor.

Black-box models are statistical models constructed using manufacturing data, and they usually have higher prediction accuracy than white-box models.[45] In contrast, they are less intuitive in nature,[45] which makes it difficult for industry operators to understand in detail. Many studies on PP-based or NIRS-based black-box models to predict the granule properties in fluidized bed granulation or drying have been reported.[4,32-34,46,47] Unlike the NIRS-based black-box models, the PP-based black-box models do not require a special initial investment because PPs are measured using standard instruments, such as a thermometer, hygrometer, flowmeter, and electric balance. Thus, researches focusing on the application of the PP-based black-box models to the fluidized bed granulation process have been conducted enthusiastically. Rambali *et al.*[4] demonstrated an optimization of the fluidized bed granulation process based on the multiple regression model. In addition, other regression models such as partial least squares regression (PLSR)[48] that can cope with collinearity among PPs have been used to increase the prediction accuracy and the robustness of the statistical model.[46,47] However, the PP-based black-box models have a certain limitation; they are valid only within the process conditions considered during model development. For example, when the manufacturing scale is changed, the prediction accuracy of the model might decrease because the relationship between the majority of PPs and the water content depends on the intrinsic properties of the equipment, such as the heat transfer coefficient. Similarly, the existing PP-based black-box models to estimate the particle size of granules were valid only for the same granulator for which experimental data was obtained.[47,49] The manufacturing scale of the granulation process is necessary to adjust according to the demand during clinical development and commercial manufacturing. In contrast, no PP-based black-box model that predicts the granule properties accurately beyond the calibrated manufacturing

scale range has been developed.

Gray-box models integrate white-box and black-box models to solve each drawback, *i.e.*, prediction accuracy and intuitiveness.[45] Liu *et al.*[50] developed a gray-box model to predict granule size distribution in fluidized bed granulation. The gray-box model integrates the PBM and PLSR model, which describes the relationship between the operating variables and kernel parameters in the PBM.[50] On the other hand, no gray-box model focusing on the granule water content in fluidized bed granulation has been reported.

## 1.4. Pharmaceutical lifecycle

According to ICH Q10[24], the pharmaceutical lifecycle is divided into four stages, *i.e.*, pharmaceutical development, technology transfer, commercial manufacturing, and product discontinuation, as summarized in Figure 1.3.

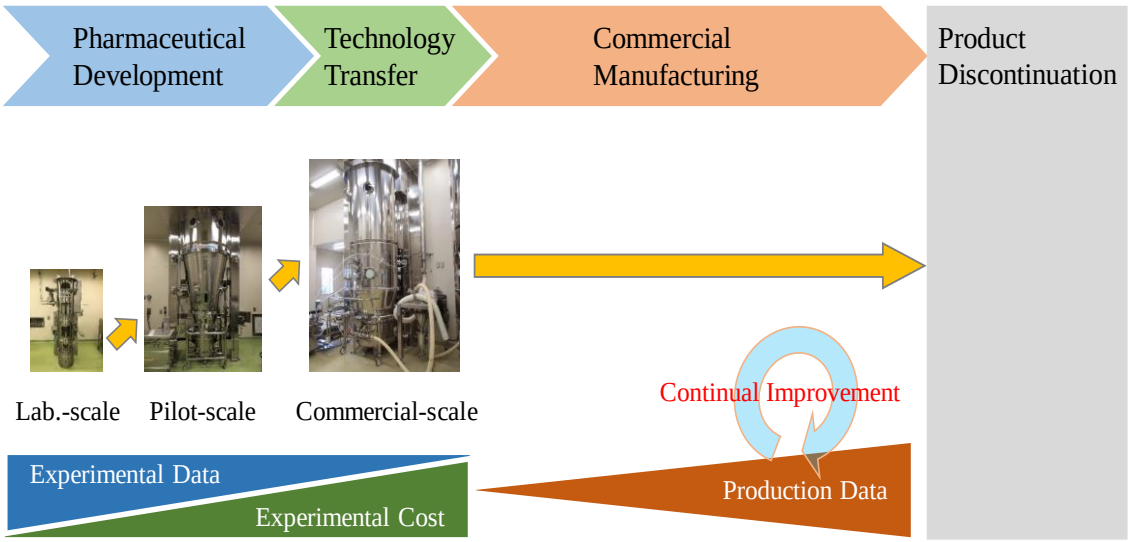


Figure 1.3. Overview of pharmaceutical lifecycle.

At the pharmaceutical development stage, formulation and manufacturing processes are developed, and investigational products are manufactured. Many experimental data are usually acquired at various operating conditions using laboratory-scale and pilot-scale equipment through this stage. Based on these experimental data, we evaluate the relationship between MAs and CQAs, which is independent of the manufacturing scale, and determine the CMAs to impact CQAs. Then, we construct a scale-free prediction model, for example, a NIRS-based black-box model, for CMA monitoring applicable to the commercial scale, using only the laboratory-scale and pilot-scale experimental data.

At the technology transfer stage, the knowledge regarding product and manufacturing processes is transferred from the development unit to the manufacturing unit through a process development study using commercial-scale equipment. In many cases, except for rare disease therapeutic drugs, the process is scaled-up in order of laboratory, pilot, and commercial scale. The process development study aims to evaluate each manufacturing process's influence on final product qualities to establish a control strategy for commercial manufacturing. However, this scaling-up is not easy because of the following two reasons. First, the number of lots in process development study using commercial-scale equipment is usually limited, for example, from three to six lots, because a massive amount of expensive API is needed. Second, the relationship between the majority of PPs and CMA differs depending on the manufacturing scale, which means that trial and error are required to clarify the setting values of PPs to satisfy the desired CMA. Hence, the scale-free prediction model for CMA monitoring is crucial for facilitating process understanding, which reveals the relationship between PPs and CMA more efficiently. An accurate NIRS-based black-box model has been widely utilized for scale-free monitoring, whereas an alternative method that requires no initial investment has been desired for

cost-saving. However, it is challenging to construct a scale-free and accurate white-box model or PP-based black-box model without using commercial-scale experimental data.

At the commercial manufacturing stage, final products with the desired quality are routinely released. In addition, continual improvement is facilitated through trend analysis and fluctuating factors identification using accumulated operating data and product knowledge, which continues until the final stage, *i.e.*, product discontinuation. A descriptive and accurate prediction model for CMA monitoring enhances industry operators' process understanding, which contributes to continual improvement at a higher level. However, a non-descriptive NIRS-based black-box model has been commonly used because a descriptive white-box model usually has a limitation of prediction accuracy. Thus, an alternative method with high prediction accuracy and interpretability is desirable.

Although the CMA monitoring method's features required at the technology transfer and the commercial manufacturing stages are different, a NIRS-based black-box model has been mainly adopted throughout the pharmaceutical lifecycle. A NIRS-based black-box model is scale-free and accurate, whereas it is non-descriptive and requires a high initial investment cost. Therefore, it is crucial to develop practical alternative CMA monitoring methods to meet all the requirements at the technology transfer (scale-free, accurate, and cost-effective) and the commercial manufacturing (descriptive, accurate, and simple) stages, respectively.

## 1.5. Aim of work and thesis outline

This thesis aims at providing practical soft sensors suitable for the technology transfer and the commercial manufacturing stages. The thesis focuses on the water content monitoring in the fluidized bed granulation considering the critical impact on final

product qualities. The following two novel soft sensors are developed.

- A scale-free, accurate, and cost-effective soft sensor for technology transfer (Chapters 3 and 4)
- A descriptive, accurate, and simple soft sensor for commercial manufacturing (Chapter 5)

In Chapter 2, the general information on this thesis, such as formulation, equipment, data acquisition, and regression methods, is described.

In Chapter 3, a scale-free, accurate, and cost-effective PP-based black-box model is proposed to address the manufacturing scale change in the technology transfer stage. This study assumed an actual technology transfer situation, where it is necessary to construct an accurate prediction model applicable to commercial scale using only laboratory-scale and pilot-scale experimental data. This study evaluates the input variables' influence on the prediction model's robustness against the manufacturing scale change to develop a scale-free statistical model. Besides, two kinds of regression methods, *i.e.*, PLSR and locally weighted partial least squares regression (LW-PLSR),[51] are evaluated to build an accurate statistical model. LW-PLSR can cope with collinearity and nonlinearity.[51]

In Chapter 4, further evaluation regarding the developed scale-free PP-based LW-PLSR model is conducted. The scale-free PP-based and NIRS-based LW-PLSR models are constructed using different calibration datasets and compared in terms of prediction accuracy and development cost. This study aims to clarify the following two points: 1) how to prepare the calibration samples required to construct accurate PP-based and NIRS-based LW-PLSR models, and 2) which type of LW-PLSR model should be selected depending on the raw material cost, composition of granules, and price of the NIR

spectrometer.

In Chapter 5, the three types of gray-box models, *i.e.*, parallel,[52] serial,[53] and combined[44] gray-box models (see Figure 1.2), are developed. The proposed gray-box models are descriptive, accurate, and simple, which means that they are valuable to the commercial manufacturing stage. In addition, as the point to consider when applying the gray-box models using LW-PLSR to commercial production, the assessment method based on Hotelling's $T^2$ and Q residual[54] is proposed to judge whether the gray-box soft sensor is suitable for precise process monitoring.

In Chapter 6, the conclusion of this thesis consisting of an overall discussion and outlook, is described.

# 2. General information

## 2.1. Formulation

The two formulations, A and B, consisted of the common API (Daiichi Sankyo Co., Ltd., Japan), whereas their components were different, as listed in Table 2.1. Besides, the mass fraction of API in formulation A was approximately twice that of formulation B. These differences in formulations A and B resulted in different water content of the blend for granulation, *i.e.*, 2.4% and 3.3%, respectively.

Table 2.1. Components of two formulations A and B.

| Components | Formulation A | Formulation B |
|---|---|---|
| API | + | + |
| Excipient A | + | + |
| Excipient B | + | + |
| Excipient C | + | + |
| Excipient D | + | − |
| Excipient E | + | − |
| Excipient F | + | − |
| Excipient G | − | + |

+: Included, −: Not included

## 2.2. Granulators

The following three fluidized bed granulators were utilized: NFLO-5 (Freund Corp., Japan) for approx. 4-kg scale, GPCG-30 (Powrex Corp., Japan) for approx. 25-kg scale, and WSG-120 (Powrex Corp.) for approx. 100-kg scale. In this thesis, the manufacturing scales were defined as follows: approx. 4 kg for the laboratory scale, approx. 25 kg for the pilot scale, and approx. 100 kg for the commercial scale. Granules were sampled throughout the granulation runs.

## 2.3. PPs measurement

The following eight PPs were measured during the granulation process: inlet air volume ($m^3$/min), inlet air temperature (°C), inlet air humidity (g-water/kg-air), spray rate (g/min), spray air volume (NL/min), product temperature (°C), exhaust air temperature (°C), and exhaust air humidity (%RH). In the eight PPs, inlet air volume ($m^3$/min), inlet air temperature (°C), spray rate (g/min), and spray air volume (NL/min) are operable PPs. All the PPs were acquired using standard instruments such as a thermometer, hygrometer, flowmeter, and electric balance. These PPs were normalized, that is, their means were equal to 0 and the variances were equal to 1.

## 2.4. Experimental data for calibration and validation

The fluidized bed granulation process is divided into two phases: the granulation phase and the drying phase. In the thesis, only the granulation phase data were utilized because the number of PPs in the granulation phase is different from that in the drying phase; spray rate (g/min) and spray air volume (NL/min) are PPs specific to the granulation phase. The granulation phase, where liquid supply and evaporation occur simultaneously, is more complex and difficult to control water content accurately than the drying phase. Besides, the scaling-up is generally performed based on the water content profile during the granulation phase. It is desirable that PPs are determined to keep the granule water content during the granulation phase constant at different manufacturing scales.[8)] Thus, in-line water content monitoring in the granulation phase is especially important, and the thesis focused on the granulation phase.

### 2.4.1. Formulation A

The experimental data utilized in Chapter 5 are summarized in Table 2.2.

### 2.4.2. Formulation B

The experimental data utilized in Chapters 3 and 4 are listed in Table 2.3, and those used in Chapter 5 are summarized in Table 2.2.

Table 2.2. Experimental data of formulations A and B using Chapter 5.

*t*: spraying time (min), PP1: inlet air volume ($m^3$/min), PP2: inlet air temperature (°C), PP3: inlet air humidity (g-water/kg-air), PP4: mean value of spray rate for 0 min to *t* min (kg/min), PP5: spray air volume (NL/min), PP6: product temperature (°C), PP7: exhaust air temperature (°C), PP8: exhaust air humidity (%RH), and WC: water content of granules (%).

| Formulation | Lot No. | Number of samples | Minimum – Maximum | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | *t* | PP1 | PP2 | PP3 | PP4 | PP5 | PP6 | PP7 | PP8 | WC |
| A | 1 | 11 | 10 – 114 | 34.6 – 37.8 | 90.1 – 90.3 | 3.5 – 3.8 | 0.747 – 0.748 | 548 – 550 | 38.4 – 43.0 | 37.8 – 42.8 | 34.8 – 52.3 | 2.6 – 3.3 |
| A | 2 | 9 | 10 – 90 | 32.7 – 33.2 | 90.1 – 90.2 | 3.5 – 3.8 | 0.948 – 0.948 | 546 – 548 | 30.2 – 36.1 | 30.2 – 38.0 | 46.9 – 90.8 | 3.5 – 15.3 |
| A | 3 | 8 | 10 – 87 | 31.7 – 33.2 | 90.2 – 90.3 | 2.8 – 3.6 | 0.948 – 0.948 | 800 – 802 | 29.8 – 35.6 | 29.7 – 37.4 | 48.4 – 91.1 | 3.9 – 13.1 |
| A | 4 | 11 | 10 – 115 | 35.8 – 38.6 | 90.1 – 90.3 | 3.4 – 3.8 | 0.747 – 0.748 | 798 – 801 | 38.4 – 43.4 | 36.9 – 43.1 | 34.4 – 53.3 | 2.4 – 3.1 |
| A | 5 | 9 | 10 – 97 | 32.6 – 35.2 | 90.2 – 90.3 | 3.6 – 3.8 | 0.847 – 0.848 | 649 – 652 | 31.4 – 39.0 | 31.3 – 40.0 | 41.4 – 83.3 | 3.1 – 6.3 |
| A | 6 | 8 | 10 – 87 | 33.8 – 35.1 | 90.1 – 90.3 | 3.6 – 3.9 | 0.947 – 0.948 | 750 – 755 | 30.2 – 37.1 | 30.2 – 38.5 | 46.8 – 90.8 | 3.7 – 12.2 |
| B | 1 | 6 | 10 – 62 | 29.9 – 30.1 | 89.9 – 90.1 | 2.2 – 2.4 | 0.698 – 0.699 | 692 – 715 | 35.5 – 41.3 | 35.1 – 40.1 | 51.1 – 69.3 | 3.6 – 5.0 |
| B | 2 | 5 | 10 – 51 | 27.4 – 28.1 | 85.0 – 85.0 | 3.8 – 4.0 | 0.848 – 0.849 | 643 – 650 | 30.1 – 35.7 | 30.8 – 37.5 | 44.7 – 77.8 | 4.3 – 10.8 |
| B | 3 | 8 | 10 – 79 | 31.2 – 32.1 | 94.9 – 95.1 | 3.8 – 4.1 | 0.548 – 0.549 | 712 – 746 | 49.8 – 51.1 | 47.0 – 49.7 | 20.4 – 26.2 | 2.5 – 2.8 |
| B | 4 | 6 | 10 – 62 | 29.3 – 30.0 | 90.0 – 90.1 | 3.6 – 4.0 | 0.697 – 0.698 | 694 – 713 | 36.5 – 42.0 | 36.0 – 41.1 | 42.1 – 62.6 | 3.5 – 4.8 |
| B | 5 | 6 | 10 – 59 | 27.6 – 28.3 | 85.0 – 85.1 | 3.7 – 4.1 | 0.749 – 0.749 | 646 – 672 | 31.1 – 38.0 | 31.5 – 38.7 | 41.9 – 74.7 | 4.1 – 8.2 |
| B | 6 | 7 | 10 – 72 | 31.4 – 32.4 | 95.0 – 95.1 | 3.7 – 4.0 | 0.598 – 0.599 | 742 – 772 | 47.1 – 47.6 | 44.9 – 45.9 | 29.0 – 31.0 | 2.7 – 3.0 |

Table 2.3. Experimental data of formulation B using Chapters 3 and 4.

*t*: spraying time (min), PP1: inlet air volume ($m^3$/min), PP2: inlet air temperature (°C), PP3: inlet air humidity (g-water/kg-air), PP4: spray rate (g/min), PP5: spray air volume (NL/min), PP6: product temperature (°C), PP7: exhaust air temperature (°C), PP8: exhaust air humidity (%RH), and WC: water content of granules (%).

| Chapter 3 | Chapter 4 | Equipment | Lot No. | *t* | PP1 | PP2 | PP3 | PP4 | PP5 | PP6 | PP7 | PP8 | WC |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Calibration | Calibration | NFLO-5 | 1 | 5 | 2.5 | 69.3 | 9.2 | 66 | 147 | 35.5 | 41.6 | 34 | 3.7 |
| Calibration | Calibration | NFLO-5 | 1 | 10 | 2.5 | 69.9 | 9.5 | 65 | 147 | 30.6 | 36.8 | 46 | 4.7 |
| Calibration | Calibration | NFLO-5 | 1 | 15 | 2.5 | 70.0 | 9.8 | 67 | 147 | 29.2 | 34.1 | 53 | 7.9 |
| Calibration | Calibration | NFLO-5 | 1 | 20 | 2.5 | 69.6 | 8.8 | 69 | 148 | 28.5 | 30.4 | 70 | 7.7 |
| Calibration | Calibration | NFLO-5 | 1 | 25 | 2.6 | 69.9 | 9.4 | 67 | 149 | 28.4 | 29.6 | 73 | 9.2 |
| Calibration | Calibration | NFLO-5 | 1 | 29 | 2.5 | 70.0 | 9.6 | 66 | 148 | 28.4 | 29.1 | 75 | 10.2 |
| Calibration | Calibration | NFLO-5 | 2 | 0 | 2.6 | 67.4 | 10.0 | 70 | 147 | 42.0 | 42.6 | 16 | 1.9 |
| Calibration | Calibration | NFLO-5 | 2 | 15 | 2.0 | 70.6 | 10.1 | 81 | 148 | 29.3 | 31.6 | 65 | 9.5 |
| Calibration | Calibration | NFLO-5 | 2 | 20 | 1.5 | 70.1 | 10.2 | 81 | 149 | 28.2 | 30.3 | 68 | 12.5 |
| Calibration | Calibration | NFLO-5 | 2 | 25 | 1.5 | 70.1 | 10.1 | 71 | 148 | 27.7 | 29.4 | 69 | 15.6 |
| Calibration | Calibration | NFLO-5 | 3 | 0 | 2.6 | 76.3 | 10.0 | 50 | 149 | 41.3 | 39.7 | 20 | 2.2 |
| Calibration | Calibration | NFLO-5 | 3 | 5 | 2.5 | 82.8 | 10.0 | 52 | 150 | 39.2 | 38.3 | 40 | 2.9 |
| Calibration | Calibration | NFLO-5 | 3 | 10 | 2.5 | 84.7 | 10.1 | 55 | 150 | 36.7 | 36.6 | 48 | 3.5 |
| Calibration | Calibration | NFLO-5 | 3 | 15 | 2.0 | 84.7 | 10.0 | 53 | 150 | 35.1 | 35.5 | 50 | 4.3 |
| Calibration | Calibration | NFLO-5 | 3 | 20 | 2.5 | 84.1 | 9.3 | 69 | 149 | 35.8 | 34.3 | 55 | 3.5 |
| Calibration | Calibration | NFLO-5 | 3 | 25 | 2.5 | 84.8 | 9.5 | 53 | 149 | 36.2 | 34.1 | 56 | 4.0 |
| Calibration | Calibration | NFLO-5 | 3 | 30 | 2.5 | 85.1 | 9.7 | 49 | 148 | 36.7 | 34.1 | 58 | 3.8 |
| Calibration | Calibration | GPCG-30 | 4 | 10 | 7.9 | 87.5 | 6.5 | 151 | 148 | 44.8 | 42.9 | 33 | 2.8 |
| Calibration | Calibration | GPCG-30 | 4 | 20 | 8.0 | 88.7 | 6.6 | 150 | 150 | 43.0 | 41.3 | 38 | 3.0 |
| Calibration | Calibration | GPCG-30 | 4 | 30 | 7.9 | 87.9 | 6.7 | 149 | 148 | 42.1 | 40.4 | 41 | 3.2 |
| Calibration | Calibration | GPCG-30 | 4 | 40 | 8.0 | 88.7 | 6.3 | 149 | 151 | 41.6 | 40.0 | 43 | 3.3 |
| Calibration | Calibration | GPCG-30 | 4 | 50 | 7.9 | 88.0 | 6.7 | 149 | 150 | 41.0 | 39.6 | 44 | 3.5 |
| Calibration | Calibration | GPCG-30 | 4 | 60 | 7.9 | 88.0 | 6.5 | 150 | 148 | 40.7 | 39.3 | 46 | 3.4 |
| Calibration | Calibration | GPCG-30 | 4 | 72 | 7.9 | 87.9 | 6.7 | 149 | 151 | 40.3 | 38.8 | 47 | 3.5 |
| Calibration | Calibration | GPCG-30 | 5 | 10 | 8.0 | 87.0 | 5.1 | 180 | 178 | 41.2 | 41.8 | 34 | 3.1 |
| Calibration | Calibration | GPCG-30 | 5 | 20 | 7.9 | 87.9 | 5.2 | 180 | 175 | 38.0 | 38.9 | 44 | 3.8 |
| Calibration | Calibration | GPCG-30 | 5 | 30 | 8.0 | 88.0 | 5.3 | 179 | 179 | 36.5 | 37.3 | 51 | 4.1 |
| Calibration | Calibration | GPCG-30 | 5 | 40 | 7.9 | 87.9 | 5.4 | 180 | 181 | 35.5 | 36.1 | 57 | 4.5 |
| Calibration | Calibration | GPCG-30 | 5 | 50 | 7.9 | 87.6 | 5.5 | 180 | 182 | 34.9 | 35.2 | 61 | 4.4 |
| Calibration | Calibration | GPCG-30 | 5 | 59 | 7.9 | 87.8 | 5.5 | 180 | 181 | 34.6 | 34.6 | 65 | 4.8 |

Table 2.3. Experimental data of formulation B using Chapters 3 and 4 (Continued).

| Chapter 3 | Chapter 4 | Equipment | Lot No. | *t* | PP1 | PP2 | PP3 | PP4 | PP5 | PP6 | PP7 | PP8 | WC |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| N/A | Calibration | WSG-120 | 6 | 10 | 29.9 | 90.1 | 3.6 | 697 | 713 | 42.0 | 41.1 | 42 | 3.5 |
| N/A | Calibration | WSG-120 | 6 | 20 | 30.0 | 90.0 | 4.0 | 697 | 695 | 39.4 | 38.9 | 49 | 4.0 |
| N/A | Calibration | WSG-120 | 6 | 30 | 29.9 | 90.1 | 3.7 | 699 | 694 | 38.0 | 37.7 | 54 | 4.4 |
| N/A | Calibration | WSG-120 | 6 | 40 | 29.9 | 90.1 | 3.8 | 698 | 703 | 37.2 | 36.9 | 58 | 4.7 |
| N/A | Calibration | WSG-120 | 6 | 50 | 30.0 | 90.1 | 3.8 | 699 | 702 | 36.7 | 36.3 | 60 | 4.7 |
| N/A | Calibration | WSG-120 | 6 | 62 | 29.3 | 90.1 | 4.0 | 699 | 704 | 36.5 | 36.0 | 63 | 4.8 |
| N/A | Calibration | WSG-120 | 7 | 10 | 31.4 | 95.0 | 3.7 | 598 | 772 | 47.4 | 44.9 | 31 | 2.8 |
| N/A | Calibration | WSG-120 | 7 | 20 | 32.4 | 95.0 | 3.7 | 598 | 752 | 47.1 | 45.0 | 30 | 2.7 |
| N/A | Calibration | WSG-120 | 7 | 30 | 32.0 | 95.1 | 4.0 | 598 | 747 | 47.2 | 45.3 | 30 | 3.0 |
| N/A | Calibration | WSG-120 | 7 | 40 | 31.5 | 95.0 | 3.8 | 600 | 747 | 47.2 | 45.5 | 30 | 3.0 |
| N/A | Calibration | WSG-120 | 7 | 50 | 31.8 | 95.0 | 3.7 | 597 | 744 | 47.3 | 45.5 | 30 | 2.9 |
| N/A | Calibration | WSG-120 | 7 | 60 | 32.0 | 95.0 | 3.7 | 598 | 746 | 47.5 | 45.7 | 29 | 2.9 |
| N/A | Calibration | WSG-120 | 7 | 72 | 31.8 | 95.0 | 3.9 | 598 | 742 | 47.6 | 45.9 | 29 | 3.0 |
| N/A | Calibration | WSG-120 | 8 | 10 | 28.1 | 85.1 | 3.9 | 749 | 672 | 38.0 | 38.7 | 42 | 4.1 |
| N/A | Calibration | WSG-120 | 8 | 20 | 28.3 | 85.1 | 3.7 | 748 | 655 | 34.4 | 35.5 | 52 | 5.3 |
| N/A | Calibration | WSG-120 | 8 | 30 | 27.6 | 85.1 | 4.0 | 750 | 650 | 32.8 | 33.7 | 61 | 6.2 |
| N/A | Calibration | WSG-120 | 8 | 40 | 28.0 | 85.0 | 3.8 | 749 | 649 | 31.9 | 32.7 | 67 | 7.1 |
| N/A | Calibration | WSG-120 | 8 | 50 | 27.9 | 85.1 | 3.9 | 749 | 648 | 31.3 | 31.9 | 72 | 7.6 |
| N/A | Calibration | WSG-120 | 8 | 59 | 27.6 | 85.0 | 4.1 | 748 | 646 | 31.1 | 31.5 | 75 | 8.2 |
| Validation | Validation | WSG-120 | 9 | 10 | 30.0 | 89.9 | 2.3 | 699 | 715 | 41.3 | 40.1 | 51 | 3.6 |
| Validation | Validation | WSG-120 | 9 | 20 | 30.1 | 90.0 | 2.2 | 699 | 692 | 38.7 | 38.1 | 54 | 4.0 |
| Validation | Validation | WSG-120 | 9 | 30 | 29.9 | 90.1 | 2.4 | 698 | 698 | 37.3 | 36.9 | 59 | 4.3 |
| Validation | Validation | WSG-120 | 9 | 40 | 30.0 | 90.0 | 2.3 | 699 | 698 | 36.5 | 36.1 | 64 | 4.7 |
| Validation | Validation | WSG-120 | 9 | 50 | 30.0 | 90.1 | 2.2 | 698 | 700 | 35.9 | 35.5 | 67 | 4.9 |
| Validation | Validation | WSG-120 | 9 | 62 | 30.0 | 90.0 | 2.3 | 697 | 699 | 35.5 | 35.1 | 69 | 5.0 |
| Validation | Validation | WSG-120 | 10 | 10 | 27.4 | 85.0 | 4.0 | 848 | 650 | 35.7 | 37.5 | 45 | 4.3 |
| Validation | Validation | WSG-120 | 10 | 20 | 28.1 | 85.0 | 3.8 | 849 | 643 | 32.2 | 33.9 | 59 | 6.2 |
| Validation | Validation | WSG-120 | 10 | 30 | 28.1 | 85.0 | 3.8 | 849 | 649 | 30.9 | 32.2 | 68 | 7.6 |
| Validation | Validation | WSG-120 | 10 | 40 | 28.1 | 85.0 | 3.9 | 848 | 646 | 30.3 | 31.2 | 74 | 9.3 |
| Validation | Validation | WSG-120 | 10 | 51 | 28.0 | 85.0 | 3.9 | 848 | 647 | 30.1 | 30.8 | 78 | 10.8 |
| Validation | Validation | WSG-120 | 11 | 10 | 31.9 | 94.9 | 3.8 | 548 | 730 | 49.8 | 47.0 | 26 | 2.6 |
| Validation | Validation | WSG-120 | 11 | 20 | 32.1 | 95.1 | 3.8 | 550 | 712 | 50.3 | 48.4 | 23 | 2.6 |
| Validation | Validation | WSG-120 | 11 | 30 | 31.5 | 95.1 | 4.1 | 548 | 746 | 50.8 | 49.2 | 22 | 2.8 |
| Validation | Validation | WSG-120 | 11 | 40 | 31.9 | 95.1 | 3.8 | 549 | 743 | 51.0 | 49.5 | 21 | 2.6 |
| Validation | Validation | WSG-120 | 11 | 50 | 32.0 | 95.0 | 3.9 | 549 | 742 | 50.9 | 49.6 | 21 | 2.8 |
| Validation | Validation | WSG-120 | 11 | 60 | 31.2 | 95.0 | 4.1 | 549 | 743 | 51.0 | 49.7 | 21 | 2.7 |
| Validation | Validation | WSG-120 | 11 | 70 | 31.8 | 95.0 | 3.9 | 549 | 741 | 51.0 | 49.7 | 21 | 2.6 |
| Validation | Validation | WSG-120 | 11 | 79 | 31.6 | 95.1 | 3.8 | 548 | 741 | 51.1 | 49.7 | 20 | 2.5 |

## 2.5. NIR spectra

### 2.5.1. Measurement

The NIR spectra of the granules were measured during the granulation process using a Fourier-transform NIR spectrometer MPA (Bruker Optik GmbH, Germany) or Matrix-F (Bruker Optik GmbH, Germany) through a fiber-optic probe attached to the fluidized bed granulator. Because the MPA and Matrix-F use an equivalent optical device, the NIR spectra obtained are comparable.

### 2.5.2. Preprocessing

The wavenumber region ranging from 8007 $cm^{-1}$ to 4258 $cm^{-1}$ was used to exclude any noisy regions, and the prediction models were constructed using the following five preprocessing methods: first derivative, second derivative, standard normal variate (SNV),[55] first derivative and SNV, and second derivative and SNV. The method presenting the highest prediction accuracy was then selected. Each preprocessing method was performed using the OPUS software (Bruker Optik GmbH, Germany).

### 2.5.3. Wavenumber selection

The wavenumber region was selected by SFD-NCSC-PLSR, which integrates spectral fluctuation dividing (SFD), nearest correlation spectral clustering (NCSC), and PLSR, to achieve high-performance prediction.[56]

## 2.6. Reference measurement

The water content of the granules was measured by the LOD device: HR73 (Mettler-

Toledo K.K., Japan) or its equivalent HR83 (Mettler-Toledo K.K.).

## 2.7. Regression method

PLSR[48] and LW-PLSR[51] were used to construct the prediction model. Recently, LW-PLSR has been applied to various industrial processes such as pharmaceutical,[49,51,57,58] petrochemical,[30,59] and semiconductor production[30,60] because of its remarkable ability to address collinearity and nonlinearity.[51] The difference between PLSR and LW-PLSR is the weighting rule for the calibration samples. In PLSR, a prediction model is built utilizing fixed weighting values for all calibration samples. In contrast, in LW-PLSR, the weighting values are determined for each query based on the distances between the query and the calibration samples, and a local PLSR model is built. Hence, LW-PLSR includes PLSR as a special case;[51] LW-PLSR is equivalent to PLSR when the localization parameter $\eta$ is set to 0, refer to Eq. (2.4) in Section 2.7.1. The PLSR and LW-PLSR models were constructed using MATLAB® software (MathWorks, Inc., US).

### 2.7.1. LW-PLSR

Input matrix $\boldsymbol{X} \in \Re^{N \times M}$ and output matrix $\boldsymbol{Y} \in \Re^{N \times L}$ are described using latent variables as follows:

$$\boldsymbol{X} = \boldsymbol{T}\boldsymbol{P}^{\mathrm{T}} + \boldsymbol{E} = \sum_{r=1}^{R} \boldsymbol{t}_r \boldsymbol{p}_r^{\mathrm{T}} + \boldsymbol{E} \tag{2.1}$$

$$\boldsymbol{Y} = \boldsymbol{T}\boldsymbol{Q}^{\mathrm{T}} + \boldsymbol{F} = \sum_{r=1}^{R} \boldsymbol{t}_r \boldsymbol{q}_r^{\mathrm{T}} + \boldsymbol{F} \tag{2.2}$$

where $\boldsymbol{T} = [\boldsymbol{t}_1\, \boldsymbol{t}_2\, \boldsymbol{t}_3 \cdots \boldsymbol{t}_R] \in \Re^{N \times R}$ is a latent variable matrix, $\boldsymbol{P} = [\boldsymbol{p}_1\, \boldsymbol{p}_2\, \boldsymbol{p}_3 \cdots \boldsymbol{p}_R] \in \Re^{M \times R}$

and $\boldsymbol{Q}=[\boldsymbol{q}_1\,\boldsymbol{q}_2\,\boldsymbol{q}_3\cdots\boldsymbol{q}_R]\in\Re^{L\times R}$ are loading matrices, $\boldsymbol{E}$ and $\boldsymbol{F}$ are errors, and $R$ is the number of latent variables.

LW-PLSR is a just-in-time modeling method. Whenever a prediction is required for a query, a local PLSR model is constructed using the weighting calibration samples. A similarity matrix to weight calibration samples is expressed as follows:

$$\boldsymbol{\Omega}=\mathrm{diag}(\omega_1\,\omega_2\,\omega_3\cdots\omega_N)\in\Re^{N\times N} \tag{2.3}$$

where similarity $\omega_n$ is calculated based on the weighted Euclidean distance $d_n$ between the query $\boldsymbol{x}_q$ and the $n$-th calibration sample $\boldsymbol{x}_n$:

$$\omega_n=\exp\left(-\frac{d_n\eta}{\sigma_d}\right) \tag{2.4}$$

$$d_n=\sqrt{(\boldsymbol{x}_n-\boldsymbol{x}_q)^{\mathrm{T}}\boldsymbol{\Lambda}(\boldsymbol{x}_n-\boldsymbol{x}_q)} \tag{2.5}$$

$$\boldsymbol{\Lambda}=\mathrm{diag}(\lambda_1\,\lambda_2\,\lambda_3\cdots\lambda_M)\in\Re^{M\times M} \tag{2.6}$$

where $\sigma_d$ is a standard deviation of $\{d_n\}$, $\eta$ is a localization parameter, $\boldsymbol{\Lambda}$ is a weighting matrix, and $\lambda_m$ is a weight of the $m$-th input variable. The tuning parameters of the LW-PLSR model, *i.e.*, $R$ (the number of latent variables) and $\eta$ (the localization parameter), are determined to minimize the prediction error sums of squares of the leave-one-out cross validation (LOOCV). The localization parameter $\eta$ was tuned in the range of 0 to 5.0 (Chapters 3 and 4) or 0 to 10.0 (Chapter 5).

The predicted output variables $\hat{\boldsymbol{y}}_q$ corresponding to the query $\boldsymbol{x}_q$ are calculated through the following steps:

I. Set $r=1$.

II. Calculate the similarity matrix $\boldsymbol{\Omega}$ through Eqs. (2.3), (2.4), (2.5), and (2.6).

III. Set $\boldsymbol{X}_r$, $\boldsymbol{Y}_r$, and $\boldsymbol{x}_{q,r}$ as follows:

$$\boldsymbol{X}_r = \boldsymbol{X} - \mathbf{1}_N \left[ \bar{x}_1 \ \bar{x}_2 \ \bar{x}_3 \cdots \bar{x}_M \right] \tag{2.7}$$

$$\boldsymbol{Y}_r = \boldsymbol{Y} - \mathbf{1}_N \left[ \bar{y}_1 \ \bar{y}_2 \ \bar{y}_3 \cdots \bar{y}_L \right] \tag{2.8}$$

$$\boldsymbol{x}_{q,r} = \boldsymbol{x}_q - \left[ \bar{x}_1 \ \bar{x}_2 \ \bar{x}_3 \cdots \bar{x}_M \right]^{\mathrm{T}} \tag{2.9}$$

$$\bar{x}_m = \frac{\sum_{n=1}^{N} \omega_n x_{nm}}{\sum_{n=1}^{N} \omega_n} \tag{2.10}$$

$$\bar{y}_l = \frac{\sum_{n=1}^{N} \omega_n y_{nl}}{\sum_{n=1}^{N} \omega_n} \tag{2.11}$$

where $\mathbf{1}_N \in \Re^N$ is a vector of ones, and $\bar{x}_m$ and $\bar{y}_l$ are weighted averages.

IV. Set $\hat{\boldsymbol{y}}_q = \left[ \bar{y}_1 \ \bar{y}_2 \ \bar{y}_3 \cdots \bar{y}_L \right]^{\mathrm{T}}$.

V. Derive the *r*-th latent variable of $\boldsymbol{X}$.

$$\boldsymbol{t}_r = \boldsymbol{X}_r \boldsymbol{w}_r \tag{2.12}$$

where the *r*-th weight $\boldsymbol{w}_r$ is derived as the eigenvector of $\boldsymbol{X}_r^{\mathrm{T}} \boldsymbol{\Omega} \boldsymbol{Y}_r \boldsymbol{Y}_r^{\mathrm{T}} \boldsymbol{\Omega} \boldsymbol{X}_r$, which corresponds to the maximum eigenvalue.

VI. Derive the *r*-th latent variable of $\boldsymbol{x}_q$.

$$t_{q,r} = \boldsymbol{x}_{q,r}^{\mathrm{T}} \boldsymbol{w}_r \tag{2.13}$$

VII. Derive the *r*-th loading vectors.

$$\boldsymbol{p}_r = \frac{\boldsymbol{X}_r^{\mathrm{T}} \boldsymbol{\Omega} \boldsymbol{t}_r}{\boldsymbol{t}_r^{\mathrm{T}} \boldsymbol{\Omega} \boldsymbol{t}_r} \tag{2.14}$$

$$\boldsymbol{q}_r = \frac{\boldsymbol{Y}_r^{\mathrm{T}} \boldsymbol{\Omega} \boldsymbol{t}_r}{\boldsymbol{t}_r^{\mathrm{T}} \boldsymbol{\Omega} \boldsymbol{t}_r} \tag{2.15}$$

VIII. Update $\hat{\boldsymbol{y}}_q \leftarrow \hat{\boldsymbol{y}}_q + t_{q,r}\boldsymbol{q}_r$.

IX. If $r = R$, end. Otherwise, set $\boldsymbol{X}_{r+1}$, $\boldsymbol{Y}_{r+1}$, and $\boldsymbol{x}_{q,r+1}$ as follows:

$$\boldsymbol{X}_{r+1} = \boldsymbol{X}_r - \boldsymbol{t}_r \boldsymbol{p}_r^{\mathrm{T}} \tag{2.16}$$

$$\boldsymbol{Y}_{r+1} = \boldsymbol{Y}_r - \boldsymbol{t}_r \boldsymbol{q}_r^{\mathrm{T}} \tag{2.17}$$

$$\boldsymbol{x}_{q,r+1} = \boldsymbol{x}_{q,r} - t_{q,r}\boldsymbol{p}_r \tag{2.18}$$

X. Set $r \leftarrow r+1$, and return to step V.

### 2.7.2. PLSR

PLSR is a linear regression method. In PLSR, a prediction model is built utilizing fixed weighting values for all calibration samples, *i.e.*, $\eta = 0$ in Eq. (2.4). The other algorithms are the same as LW-PLSR. The tuning parameter of the PLSR model, *i.e.*, $R$ (the number of latent variables), is determined to minimize the prediction error sums of squares of the LOOCV.

## 2.8. Prediction performance evaluation

To evaluate the prediction accuracy of the developed models, the root mean square error of calibration ($RMSE_C$), the root mean square error of cross validation ($RMSE_{CV}$), the root mean square error of prediction ($RMSE_P$), the correlation coefficient (R), and coefficient of determination ($R^2$) were calculated.

RMSE, R, and $R^2$ are defined as:

$$\mathrm{RMSE} = \sqrt{\frac{1}{K}\sum_{k=1}^{K}\left(y_k - \hat{y}_k\right)^2} \tag{2.19}$$

where $K$ is the number of samples, and $y_k$ and $\hat{y}_k$ are the reference value and prediction value of the water content for the $k$-th sample, respectively.

$$R = \frac{\sum_{k=1}^{K}\left(y_k - \bar{y}\right)\left(\hat{y}_k - \bar{\hat{y}}\right)}{\sqrt{\sum_{k=1}^{K}\left(y_k - \bar{y}\right)^2}\sqrt{\sum_{k=1}^{K}\left(\hat{y}_k - \bar{\hat{y}}\right)^2}} \tag{2.20}$$

where $\bar{y}$ and $\bar{\hat{y}}$ are the mean values of $y_k$ and $\hat{y}_k$, respectively.

$$R^2 = 1 - \frac{\sum_{k=1}^{K}\left(y_k - \hat{y}_k\right)^2}{\sum_{k=1}^{K}\left(y_k - \bar{y}\right)^2} \tag{2.21}$$

## 2.9. Evaluation of LW-PLSR model

Two distance criteria, *i.e.*, Hotelling's $T^2$ and Q residual,[54] were utilized to assess the validity of the LW-PLSR model for a query. Kamohara *et al.*[61] and Muteki *et al.*[62] reported that $T^2$ and Q could test the validity of the PLSR model for the query. In this thesis, a 99% confidence limit was adopted as the threshold of $T^2$ and Q to test whether the LW-PLSR model was valid for the query. $T^2$ and Q values were calculated using the following equations:

$$T^2 = \sum_{r=1}^{R}\frac{\left(t_{r,query} - \bar{t}_{r,calibration}\right)^2}{s^2_{t_{r,calibration}}} \tag{2.22}$$

$$Q = \sum_{m=1}^{M}\left(x_{m,query} - \hat{x}_{m,query}\right)^2 \tag{2.23}$$

where $R$ is the number of latent variables, $t_{r,query}$ is the $r$-th latent variable score of the query, $\bar{t}_{r,calibration}$ is the mean value of $\boldsymbol{t}_{r,calibration}$, which is the $r$-th score vector of the

calibration dataset samples, $s^2_{\boldsymbol{t}_{r,calibration}}$ is the variance of $\boldsymbol{t}_{r,calibration}$, $M$ is the number of input variables, $x_{m,query}$ and $\hat{x}_{m,query}$ are the query's experimental value and reconstructed value of the $m$-th input variable.

# 3. Scale-free PP-based Black-box Soft Sensor

## 3.1. Introduction

This study aims to develop an accurate PP-based black-box model that can counter the manufacturing scale change. This study evaluated the effects of the input variables and a regression method on the accuracy and robustness of the prediction model. There are two key ideas to build a scale-free PP-based black-box model for water content monitoring. First, to accommodate the manufacturing scale change, the input variables were selected based on variable importance in the PLSR model constructed using the experimental data obtained at different manufacturing scales. Second, to construct an accurate statistical model, LW-PLSR was utilized. The LW-PLSR is a type of just-in-time modeling method that can cope with collinearity and nonlinearity.[51] The LW-PLSR was adopted because the relationship between basic PPs and water content could be nonlinear. In this study, PP-based black-box models were developed using both laboratory-scale and pilot-scale experimental data; the prediction accuracy in the commercial scale was evaluated based on the assumption that the process was scaled-up from the pilot scale to the commercial scale. A NIRS-based prediction model was also constructed as a reference. Both the PP-based and NIRS-based black-box models were constructed using a dataset obtained from the same experiments, and their prediction accuracy was compared.

## 3.2. Materials and Methods

### 3.2.1. Materials

The formulation B was granulated using laboratory, pilot, and commercial-scale fluidized bed granulators, refer to Sections 2.1 and 2.2.

### 3.2.2. Measurement of PPs, NIR spectra, and reference

Refer to Sections 2.3, 2.5.1, and 2.6.

### 3.2.3. Calibration and validation datasets

In this study, a scenario in which the granulation process was scaled-up from a pilot scale to a commercial scale was investigated. As shown in Table 2.3, the calibration dataset for both the PP-based and NIRS-based models consists of 30 samples (Lot Nos. 1-5), which were acquired at the laboratory and pilot scale: NFLO-5 (3 lots) and GPCG-30 (2 lots). On the other hand, the validation dataset consists of 19 samples (Lot Nos. 9-11) acquired at the commercial scale: WSG-120 (3 lots), which is not included in the calibration dataset.

### 3.2.4. Input variable selection

#### 3.2.4.1. PPs

To predict the water content at a commercial scale without using the commercial-scale experimental data in the calibration dataset, critical PPs whose effects on water content are constant among manufacturing scales were selected as follows. First, a PLSR model was constructed using all the PPs provided in the calibration dataset. Then, variable importance in the projection (VIP) scores, which indicate contributions of the input variables to the prediction value, were calculated. The VIP score for the *j*-th variable is defined as follows:

$$VIP_j = \sqrt{\frac{M \sum_{r=1}^{R} \left[ \left( q_r^2 t_r^T t_r \right) \left( w_{jr} / \| w_r \| \right)^2 \right]}{\sum_{r=1}^{R} \left( q_r^2 t_r^T t_r \right)}} \quad (3.1)$$

where $M$ is the number of input variables, $R$ is the number of latent variables, $\boldsymbol{q}_r$ is the $r$-th regression coefficient vector, $\boldsymbol{t}_r$ is the $r$-th score vector, and $w_{jr}$ is the $j$-th element of the $r$-th weighting vector $\boldsymbol{w}_r$. Here, $R$ was determined to minimize the prediction error sums of squares of the LOOCV.

In general, the input variables whose VIP scores are greater than or equal to one are considered critical to the prediction value.[63] In this study, PPs with VIP scores greater than or equal to one were selected as the input variables.

#### 3.2.4.2. NIR spectra

Refer to Sections 2.5.2, and 2.5.3.

### 3.2.5. Regression method

PLSR and LW-PLSR were utilized to build the prediction model using the selected input variables. To evaluate the effect of the input variable selection on the prediction performance, a PLSR model using all the PPs was constructed and evaluated. LOOCV was performed to determine the tuning parameters of PLSR and LW-PLSR models. The detailed information regarding PLSR and LW-PLSR is described in Section 2.7.

### 3.2.6. Prediction performance evaluation

After the LOOCV was conducted, $RMSE_{CV}$ and $R_{CV}$ were calculated. Besides, to evaluate the applicability to the validation dataset, $RMSE_P$ and $R_P$ were calculated. The detailed information on calculation is described in Section 2.8.

### 3.2.7. Evaluation of LW-PLSR model

Refer to Section 2.9.

## 3.3. Results and discussion

### 3.3.1. PP-based black-box models

#### 3.3.1.1. Input variable selection

Figure 3.1 shows the VIP scores of all the PPs, which were calculated using the PLSR model after normalization. The results show that the VIP scores of four PPs, *i.e.*, inlet air temperature (°C), product temperature (°C), exhaust air temperature (°C), and exhaust air humidity (%RH), were greater than one. Therefore, these four PPs were selected as the input variables.

$RMSE_P$ values of the PLSR models are provided in Figures 3.2A and 3.2B. $RMSE_P$ was significantly reduced from 27.2% to 3.5% by using the four selected PPs as input variables. In addition, $RMSE_{CV}$ values are provided in Figures 3.2C and 3.2D. The PLSR model with the four selected PPs showed a notably smaller difference between $RMSE_P$ and $RMSE_{CV}$ than that with all PPs; their differences were 1.7% and 25.6%, respectively. These results confirm that the proposed method constructed a robust PP-based black-box model to counter the manufacturing scale change.

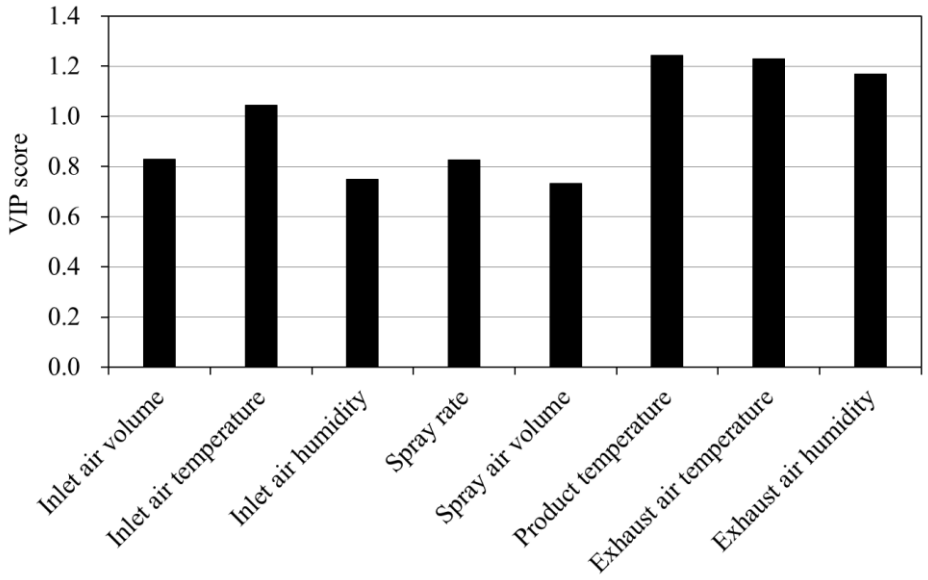


Figure 3.1. Importance of PPs based on VIP scores.

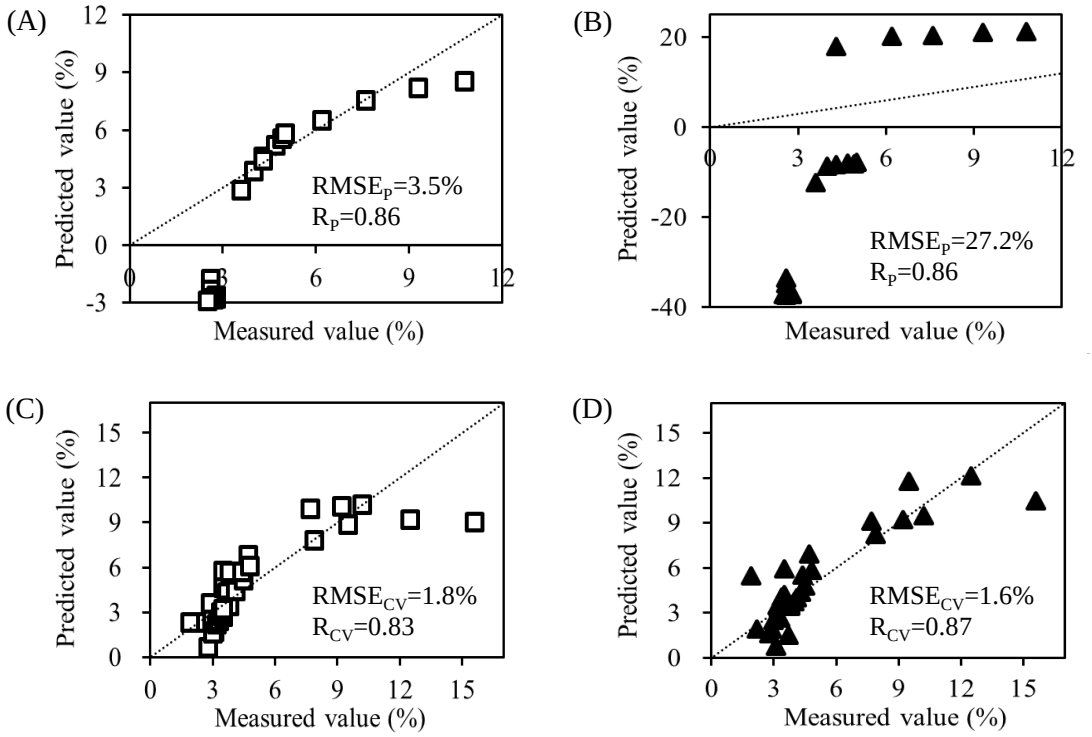


Figure 3.2. Prediction accuracy of PLSR models using the four selected PPs (A, C) and all PPs (B, D).

### 3.3.1.2. Regression method

LW-PLSR and PLSR were compared in terms of prediction accuracy. The four selected PPs were used as input variables in both models; $RMSE_P$ values were calculated using the validation dataset shown in Table 2.3. As shown in Figures 3.2A and 3.3A, the $RMSE_P$ of 0.5% for the LW-PLSR model was significantly smaller than that of 3.5% for the PLSR model. In addition, the difference between the $RMSE_P$ and $RMSE_{CV}$ of -0.2% for the LW-PLSR model was smaller than that of 1.7% for the PLSR model, as shown in Figures 3.2 and 3.3. These results indicate that the LW-PLSR model is more accurate and robust than the PLSR model. The LW-PLSR model with the four selected PPs was confirmed to show the highest prediction accuracy among the three PP-based black-box models. LW-PLSR was effective in improving the prediction accuracy because the relationship between the four selected input variables and the output variable of the water content was nonlinear, as shown in Figure 3.4. Hence, the LW-PLSR model achieved a better prediction performance and robustness compared to the PLSR model.

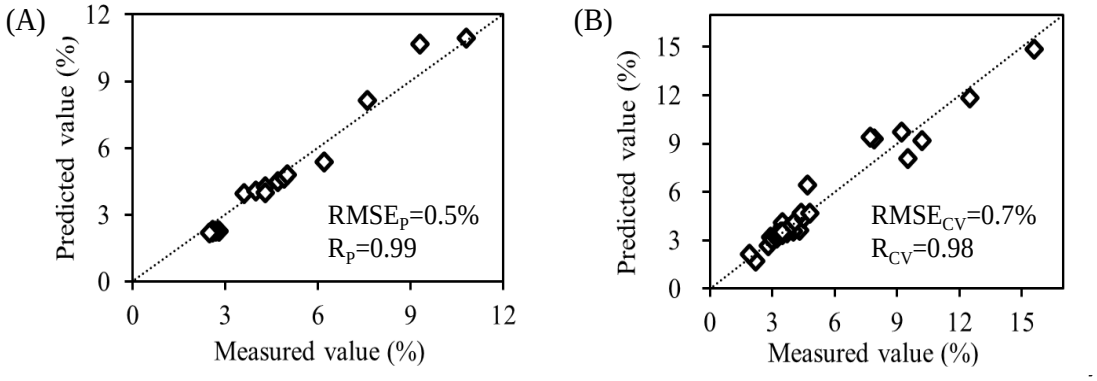


Figure 3.3. Prediction accuracy of LW-PLSR model using the four selected PPs.

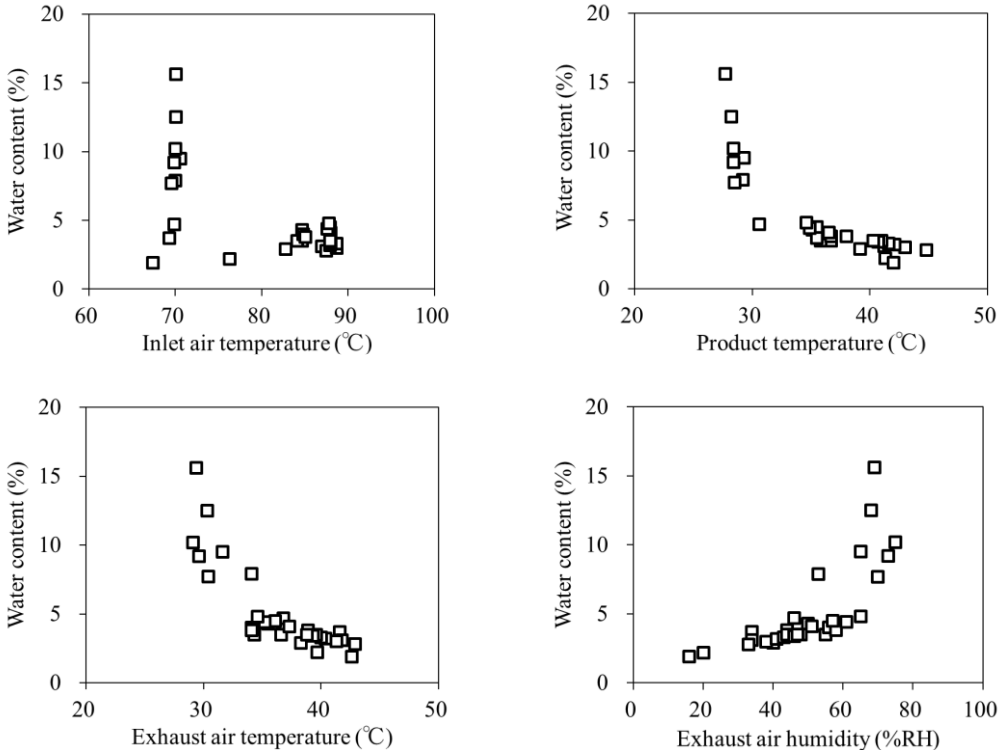


Figure 3.4. Relationship between input and output variables in the calibration dataset.

As shown in Figure 3.5, the four PPs selected as input variables, *i.e.*, inlet air temperature (°C), product temperature (°C), exhaust air temperature (°C), and exhaust air humidity (%RH), were in similar ranges regardless of the manufacturing scale. In contrast, the parameter ranges of the other four PPs, *i.e.*, inlet air volume ($m^3$/min), inlet air humidity (g-water/kg-air), spray rate (g/min), and spray air volume (NL/min), significantly differed depending on the manufacturing scale. Reflecting the ranges of the four PPs used in the LW-PLSR model, $T^2$ and Q values of all validation dataset samples were within the 99% confidence limit, which contributes to the validity of the LW-PLSR models for the query.

Furthermore, the reasonability of the VIP criteria was retrospectively evaluated. The LW-PLSR model utilizing the four PPs showed higher prediction accuracy compared to

those using different PPs selected by the various VIP criteria, as shown in Figure 3.6. The difference between the $RMSE_P$ and $RMSE_{CV}$ became remarkably smaller when the VIP criterion was 1.0 or greater. Additionally, the $RMSE_P$ and $RMSE_{CV}$ values of the LW-PLSR model using the four PPs selected based on the VIP criterion of 1.0, were smaller than those utilizing the two PPs selected at the VIP criterion of 1.2. Based on the results, the four selected PPs were considered critical for constructing a robust PP-based black-box model to counter the manufacturing scale change.

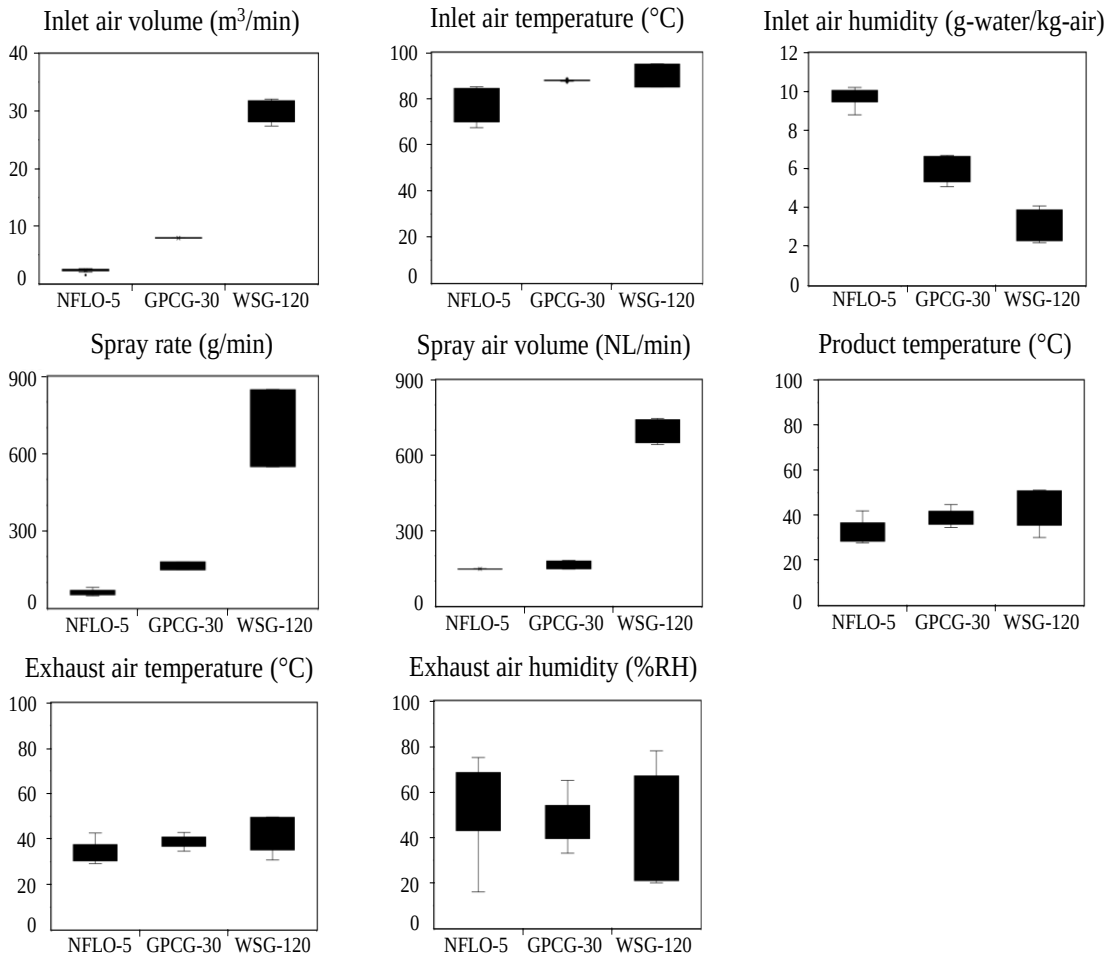


Figure 3.5. Input variable ranges at three fluidized bed granulators with different scales.

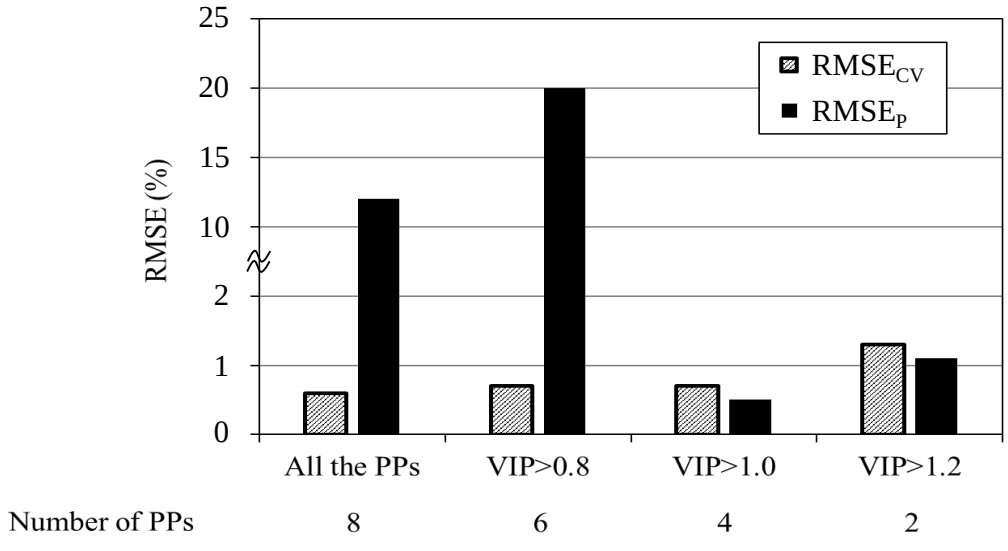


Figure 3.6. Difference between $RMSE_P$ and $RMSE_{CV}$ values of LW-PLSR models depending on VIP score criteria.

### 3.3.2. NIRS-based black-box models

#### 3.3.2.1. Preprocessing

Table 3.1 shows the $RMSE_{CV}$ and $R_{CV}$ values of the five PLSR models constructed by applying various preprocessing methods. As a result, SNV, which showed the highest prediction accuracy, was selected as the preprocessing method for constructing the NIRS-based black-box model.

Table 3.1. Comparison of five preprocessing methods in $RMSE_{CV}$ and $R_{CV}$ values.

| Preprocessing method | $RMSE_{CV}$ (%) | $R_{CV}$ |
|---|---|---|
| First derivative | 0.6 | 0.98 |
| Second derivative | 0.7 | 0.97 |
| SNV | 0.5 | 0.99 |
| First derivative + SNV | 0.7 | 0.98 |
| Second derivative + SNV | 0.8 | 0.97 |

### 3.3.2.2. Wavenumber selection

Figure 3.7 shows the wavenumber region selected by SFD-NCSC-PLSR, *i.e.*, ranging from 7212 $cm^{-1}$ to 6935 $cm^{-1}$, which includes the first overtone wavenumber region for $H_2O$.

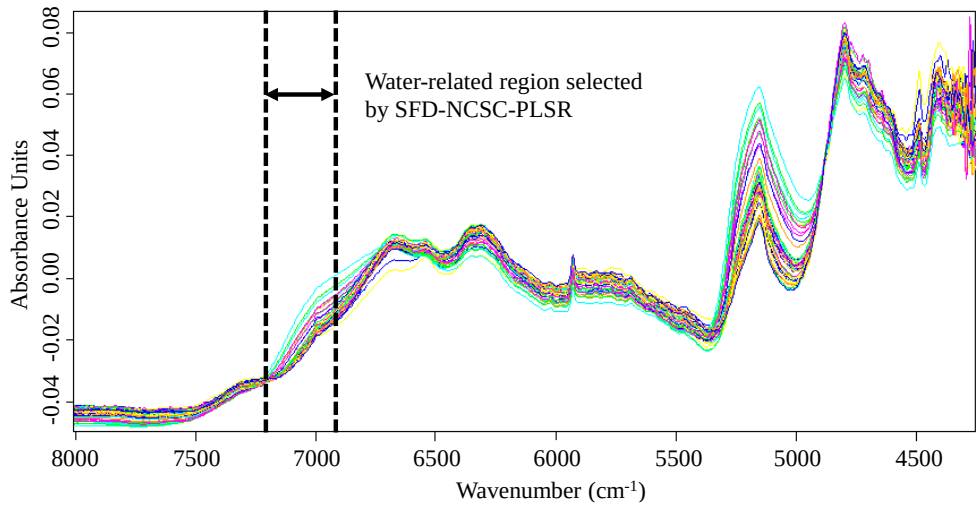


Figure 3.7. NIR spectra, preprocessed using SNV, of granules during fluidized bed granulation at the wavenumber region ranging from 8007 $cm^{-1}$ to 4258 $cm^{-1}$.

### 3.3.2.3. Regression method

The LW-PLSR and PLSR models were constructed after applying SNV to the calibration dataset of the NIR spectra with a wavenumber region ranging from 7212 $cm^{-1}$ to 6935 $cm^{-1}$. The $R_P$ values for both models were evaluated using the validation dataset shown in Table 2.3. The $R_P$ value of 0.98 for the LW-PLSR model was higher than that of 0.93 for the PLSR model, as observed from the results shown in Figures 3.8A and 3.8B. Moreover, Figure 3.8 shows the 0.5% difference between the $RMSE_P$ and $RMSE_{CV}$ for the LW-PLSR model, which was equivalent to that of 0.6% for the PLSR model. Thus,

the NIRS-based LW-PLSR model was confirmed to be the better estimator of the water content compared to the PLSR model.

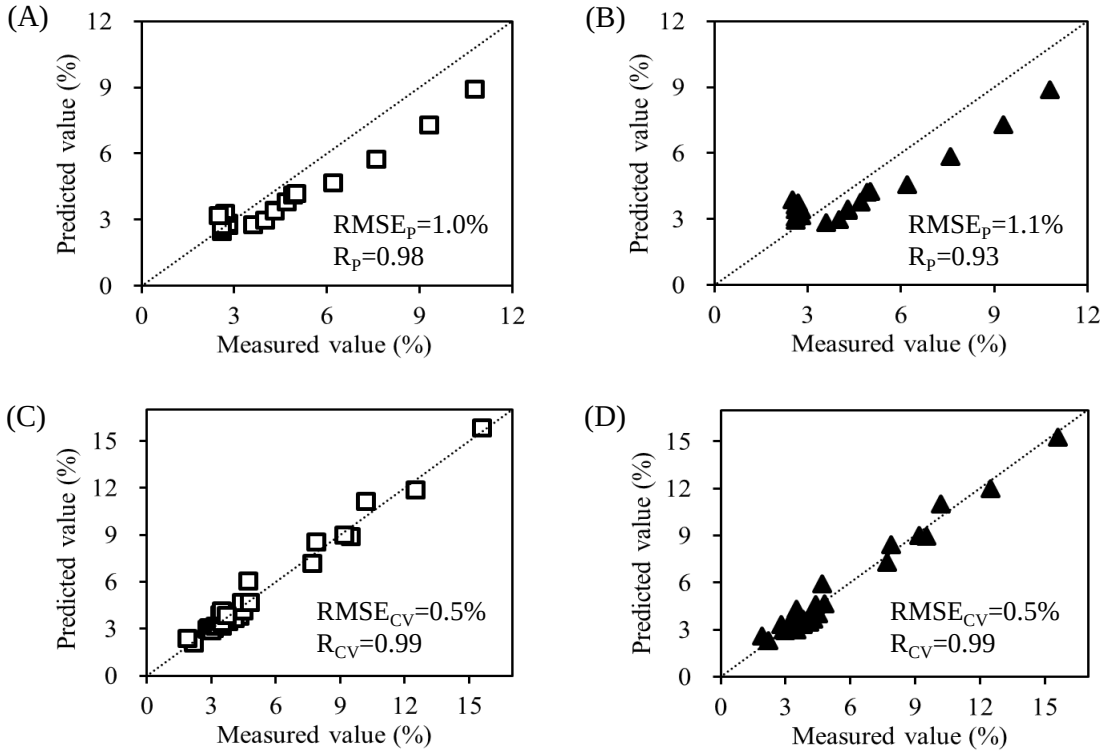


Figure 3.8. Prediction accuracy of LW-PLSR (A, C) and PLSR (B, D) models using absorbance at the wavelength selected by SFD-NCSC-PLSR.

### 3.3.3. Comparison of the prediction accuracy

The best PP-based LW-PLSR model with the four PPs, and the best NIRS-based LW-PLSR model with a wavenumber region from 7212 $cm^{-1}$ to 6935 $cm^{-1}$ were compared. As shown in Figures 3.3A and 3.8A, the 0.5% $RMSE_P$ for the PP-based LW-PLSR model was slightly smaller than that of 1.0% for the NIRS-based LW-PLSR model. In addition, Figures 3.3A, 3.3B, 3.8A, and 3.8C show the -0.2% difference between the $RMSE_P$ and $RMSE_{CV}$ for the PP-based LW-PLSR model, which was equivalent to that of 0.5% for the

NIRS-based LW-PLSR model. These results indicate that the PP-based LW-PLSR model was as accurate as the NIRS-based LW-PLSR model and robust enough to counter the manufacturing scale change.

## 3.4. Summary

In Chapter 3, it was demonstrated that the proposed PP-based black-box model could predict water content accurately beyond the manufacturing scale range of the calibration dataset. There are two key ideas for constructing a scale-free PP-based black-box model. First, to accommodate the manufacturing scale change, critical PPs whose effects on water content are constant among manufacturing scales were used as the input variables. Second, to construct an accurate statistical model, LW-PLSR was utilized to cope with collinearity and nonlinearity. The experiments demonstrated that the PP-based black-box model based on the proposed method exhibited a high prediction accuracy, which was equivalent to the widely-used NIRS-based black-box model. Unlike the NIRS-based method, the proposed method requires only standard instruments. Therefore, the proposed method is expected to be a cost-effective alternative to the existing NIRS-based monitoring method.

# 4. Prediction Performance and Economic Efficiency of Black-box Soft Sensors: PP-based Model vs. NIRS-based Model

## 4.1. Introduction

PP-based and NIRS-based models are standard statistical models used to predict water content. Economic efficiency, as well as prediction accuracy, is a key factor to be considered when choosing between PP-based models and NIRS-based models. Cogdill *et al.*[64] demonstrated through a hypothetical case study that investment in PAT generated financial benefit through the achievement of real-time release testing (RTRT), which reduces the labor and resources required for the operation of the quality unit. Thus, the economic potential of PAT in a routine production period was revealed. However, an economic case study focusing on the development cost of PAT has not been reported. To maximize the financial benefit of PAT, it is crucial to reduce its development cost.

In Chapter 3, the application of the scale-free PP-based LW-PLSR model for water content monitoring was demonstrated. The PP-based and NIRS-based LW-PLSR models, which were constructed using a common calibration dataset, showed equivalent prediction accuracy. However, the influence of the calibration dataset on the prediction accuracy has not been revealed; therefore, further evaluation is needed to understand which properties of the dataset are crucial to building accurate scale-free PP-based LW-PLSR models.

In this study, two types of statistical models to predict water content, *i.e.*, PP-based and NIRS-based LW-PLSR models, were constructed using different calibration datasets and compared in terms of prediction accuracy and development cost. This study aims to reveal the following two points: 1) how to prepare the calibration samples required to

construct accurate PP-based and NIRS-based LW-PLSR models, and 2) which type of statistical model should be selected depending on the raw material cost, composition of granules, and price of the NIR spectrometer. Clarification of these two points enables to promote the efficiency of building accurate models and save the development cost of PAT, which in turn improves the financial benefit of PAT and enhances more accurate quality control.

## 4.2. Materials and Methods

PP-based and NIRS-based LW-PLSR models were developed, and their prediction accuracy and development cost were compared. The development cost of PAT was defined as the cost of goods required to construct an accurate model of commercial-scale equipment. Various calibration datasets were prepared in accordance with the general process development scenario; that is, the process was scaled-up in order of laboratory, pilot, and commercial scale. To evaluate the development cost comprehensively, the raw material cost, composition of granules, and price of the NIR spectrometer were taken into account.

### 4.2.1. Materials

The formulation B was granulated using laboratory, pilot, and commercial-scale fluidized bed granulators, refer to Sections 2.1 and 2.2.

### 4.2.2. Measurement of PPs, NIR spectra, and reference

Refer to Sections 2.3, 2.5.1, and 2.6.

### 4.2.3. Calibration and validation datasets

As shown in Table 2.3, the eleven granulation lot data (Lot Nos. 1-11) were utilized. Three commercial-scale granulation lots (Lot Nos. 9-11) were selected from the eleven lots as a validation dataset. From the remaining eight granulation lots (Lot Nos. 1-8), ten calibration datasets (calibration dataset A to J) were prepared, as shown in Table 4.1. The number of granulation lots in the calibration datasets was increased from one (calibration dataset A) to eight (calibration dataset H), considering the general process development scenario; that is, the process is scaled-up in order of laboratory, pilot, and commercial scale. Additionally, calibration datasets I and J consisted of only pilot-scale and commercial-scale experimental data, respectively. PP-based and NIRS-based models were developed using each calibration dataset.

Table 4.1. Ten calibration datasets consisting of different manufacturing scale data.

| Calibration dataset | Lot No. | | | Number of samples | | |
|---|---|---|---|---|---|---|
| | Laboratory | Pilot | Commercial | Laboratory | Pilot | Commercial |
| A | 1 | - | - | 6 | 0 | 0 |
| B | 1, 2 | - | - | 10 | 0 | 0 |
| C | 1, 2, 3 | - | - | 17 | 0 | 0 |
| D | 1, 2, 3 | 4 | - | 17 | 7 | 0 |
| E | 1, 2, 3 | 4, 5 | - | 17 | 13 | 0 |
| F | 1, 2, 3 | 4, 5 | 6 | 17 | 13 | 6 |
| G | 1, 2, 3 | 4, 5 | 6, 7 | 17 | 13 | 13 |
| H | 1, 2, 3 | 4, 5 | 6, 7, 8 | 17 | 13 | 19 |
| I | - | 4, 5 | - | 0 | 13 | 0 |
| J | - | - | 6, 7, 8 | 0 | 0 | 19 |

Calibration datasets A, B, and C consisted of one or more laboratory-scale lots, *i.e.*, Lot Nos. 1, 2, and 3. Calibration datasets D and E consisted of three laboratory-scale lots and one or two pilot-scale lots, *i.e.*, Lot Nos. 4 and 5. Calibration datasets F, G, and H consisted of three laboratory-scale lots, two pilot-scale lots, and one or more commercial-scale lots, *i.e.*, Lot Nos. 6, 7, and 8. Calibration dataset I consisted of only two pilot-scale

lots, *i.e.*, Lot Nos. 4 and 5. Finally, calibration dataset J consisted of only three commercial-scale lots, *i.e.*, Lot Nos. 6, 7, and 8. The validation dataset consisted of three commercial-scale lots, *i.e.*, Lot Nos. 9, 10, and 11, to judge whether the constructed models could predict water content accurately on a commercial scale.

To assess the similarity between calibration and validation samples, principal component analysis (PCA)[65] was applied to the calibration dataset H. The validation samples were projected onto the subspace spanned by the first and second principal components (PCs). PCA models were constructed using Python software (Python Software Foundation).

### 4.2.4. Input variables

#### 4.2.4.1. PP-based models

From the eight PPs provided in Table 2.3, the following four critical PPs whose effects on water content are constant among manufacturing scales were utilized, based on the result in Chapter 3 (see Section 3.3.1.1): inlet air temperature (°C), product temperature (°C), exhaust air temperature (°C), and exhaust air humidity (%RH).

#### 4.2.4.2. NIRS-based models

This study used SNV as the preprocessing method and adopted the wavenumber region from 7212 $cm^{-1}$ to 6935 $cm^{-1}$ selected by SFD-NCSC-PLSR, based on the result in Chapter 3 (see Sections 3.3.2.1 and 3.3.2.2).

### 4.2.5. Regression method

In both PP-based and NIRS-based models, it was demonstrated that LW-PLSR models were the better estimators of water content than the PLSR models, based on the result in Chapter 3 (see Sections 3.3.1.2 and 3.3.2.3). Hence, this study utilized LW-PLSR was used to construct both PP-based and NIRS-based models. The detailed information on LW-PLSR is described in Section 2.7.1.

### 4.2.6. Prediction performance evaluation

To evaluate the applicability of water content monitoring on commercial scale, $RMSE_P$ of the constructed models was evaluated. The detailed information on calculation is described in Section 2.8.

### 4.2.7. Development cost of PAT

The development cost of PAT ($) was defined as the cost of goods required to construct an accurate model on the commercial scale. The required criterion of $RMSE_P$ was set to 1.0% because Alcalà *et al.*[34] reported that the water content prediction model, whose $RMSE_P$ was 0.93%, was useful for monitoring the granulation process.

The development cost of PAT $Cost_{dev}$ is expressed as:

$$Cost_{dev} = Cost_{rm} + Cost_{inv} \quad (4.1)$$

where $Cost_{rm}$ is the raw material cost ($) to obtain the experimental data required to build an accurate model, whose $RMSE_P$ is smaller than 1.0%, and $Cost_{inv}$ is the initial investment cost ($), which is the price of the NIR spectrometer. $Cost_{inv}$ was regarded

as a variable; it was set as $\$2.0\times10^4$, $\$1.1\times10^5$, and $\$2.0\times10^5$ considering the actual price of NIR spectrometers.

The raw material cost $Cost_{rm}$ was calculated by

$$Cost_{rm} = Cost_{rm,lab} + Cost_{rm,pilot} + Cost_{rm,coml} \tag{4.2}$$

where $Cost_{rm,lab}$, $Cost_{rm,pilot}$, and $Cost_{rm,coml}$ are the raw material cost (\$) to obtain the experimental data at the laboratory, pilot, and commercial scale, respectively. Then, $Cost_{rm,lab}$, $Cost_{rm,pilot}$, and $Cost_{rm,coml}$ are defined as follows:

$$Cost_{rm,lab} = \left[ Cost_{API} G_{lab} MF_{API} + Cost_{ex} G_{lab} \left(1 - MF_{API}\right) \right] N_{lab} \tag{4.3}$$

$$Cost_{rm,pilot} = \left[ Cost_{API} G_{pilot} MF_{API} + Cost_{ex} G_{pilot} \left(1 - MF_{API}\right) \right] N_{pilot} \tag{4.4}$$

$$Cost_{rm,coml} = \left[ Cost_{API} G_{coml} MF_{API} + Cost_{ex} G_{coml} \left(1 - MF_{API}\right) \right] N_{coml} \tag{4.5}$$

where $Cost_{API}$ is the cost of API (\$/kg), $G_{lab}$, $G_{pilot}$, and $G_{coml}$ are the mass of the powders charged into the granulator at laboratory, pilot, and commercial scale (kg), $MF_{API}$ is the mass fraction of API, $Cost_{ex}$ is the cost of excipients (\$/kg), and $N_{lab}$, $N_{pilot}$, and $N_{coml}$ are the minimum required numbers of the experimental lots at the laboratory, pilot, and commercial scale, which are necessary to construct an accurate model. In this study, to perform an exhaustive economic evaluation, $Cost_{API}$ and $MF_{API}$ were set as variables; $Cost_{API}$ varied from 0 \$/kg to $2.0\times10^4$ \$/kg, and $MF_{API}$ varied from 0 to 1. $Cost_{ex}$ was fixed at 40 \$/kg because it is usually much cheaper than $Cost_{API}$. $G_{lab}$, $G_{pilot}$, and $G_{coml}$ were set as 4.3 kg, 24.0 kg, and 96.2 kg, respectively.

To compare the economic efficiency of the PP-based model and the NIRS-based model, the difference in development cost was calculated by

$$Cost_{difference} = Cost_{dev,PP} - Cost_{dev,NIRS} \tag{4.6}$$

where $Cost_{dev,PP}$ is the development cost of the PP-based model, and $Cost_{dev,NIRS}$ is that of the NIRS-based model.

## 4.3. Results and discussion

### 4.3.1. Prediction accuracy

#### 4.3.1.1. PP-based LW-PLSR models

Figure 4.1 shows the $RMSE_P$ values of the PP-based LW-PLSR models, constructed utilizing the four PPs: inlet air temperature (°C), product temperature (°C), exhaust air temperature (°C), and exhaust air humidity (%RH). The $RMSE_P$ values of LW-PLSR models built using experimental data acquired at only the laboratory scale, *i.e.*, calibration datasets A, B, and C, were larger than 1.0. In contrast, the LW-PLSR model constructed utilizing the three laboratory-scale lots and one pilot-scale lot (calibration dataset D), met the criterion of $RMSE_P$. Thus, $N_{lab}$, $N_{pilot}$, and $N_{coml}$ were determined to be three, one, and zero, respectively.

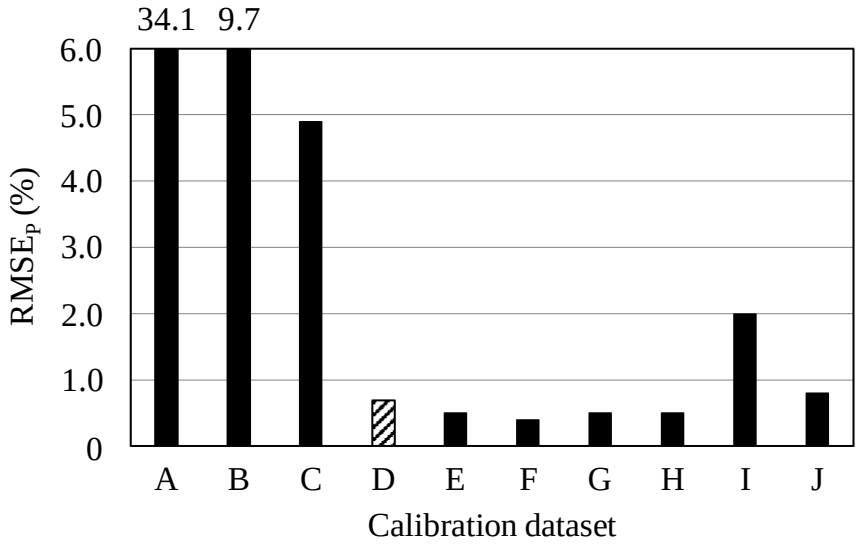


Figure 4.1. Prediction accuracy of PP-based LW-PLSR models.

The diagonal bar indicates the calibration dataset with the minimum required number of experimental data to meet the criterion of $RMSE_P$.

PCA was applied to the calibration dataset H. The relationship between the number of PCs and the cumulative proportion is shown in Figure 4.2. With the first two PCs, the cumulative proportion surpassed 0.9. The scatter plots of the first and second PCs are provided in Figure 4.3. Calibration samples, Lot Nos. 1 and 2, had a considerable distance from the validation samples compared with the other calibration samples. As shown in Table 2.3, the inlet air temperatures in Lot Nos. 1 and 2 were around 70 °C, which was much lower than 85-95 °C in the validation samples, *i.e.*, Lot Nos. 9, 10, and 11. According to the loading scores, as shown in Figure 4.4, the inlet air temperature had an impact on both the first and second PCs. Reflecting the distance between calibration samples and validation samples on the PC1-PC2 subspace, the $RMSE_P$ values of 34.1% and 9.7% for the models constructed using datasets A and B considerably exceeded the criterion of 1.0%. On the other hand, the $RMSE_P$ values of 4.9% and 0.7% for the models

constructed using datasets C and D were remarkably small because the calibration samples, Lot Nos. 3 and 4, were located near the validation samples on the PC1-PC2 subspace. The $RMSE_P$ of 0.7% for the model built by dataset D was equivalent to $RMSE_P$ values of 0.5%, 0.4%, 0.5%, and 0.5% for the models constructed using datasets E, F, G, and H, respectively. Thus, dataset D contained sufficient data to construct an accurate model, and additional data did not improve prediction performance. Also, the $RMSE_P$ of 2.0% for the model constructed by dataset I was smaller than that of 4.9% for the model built by dataset C because the calibration samples consisting of dataset I were closer to the validation samples than those consisting of dataset C on the PC1-PC2 subspace. As shown in Table 4.1, dataset I consisted of two pilot-scale lots, while dataset C contained three laboratory-scale lots. These results indicated that the prediction accuracy of PP-based LW-PLSR models was dependent on the distance between calibration and validation samples on the PC1-PC2 subspace.

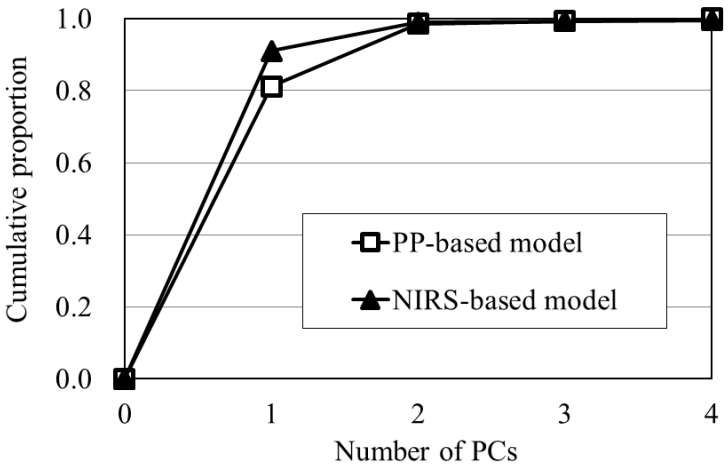


Figure 4.2. Relationship between number of PCs and cumulative proportion in PCA conducted using calibration dataset H.

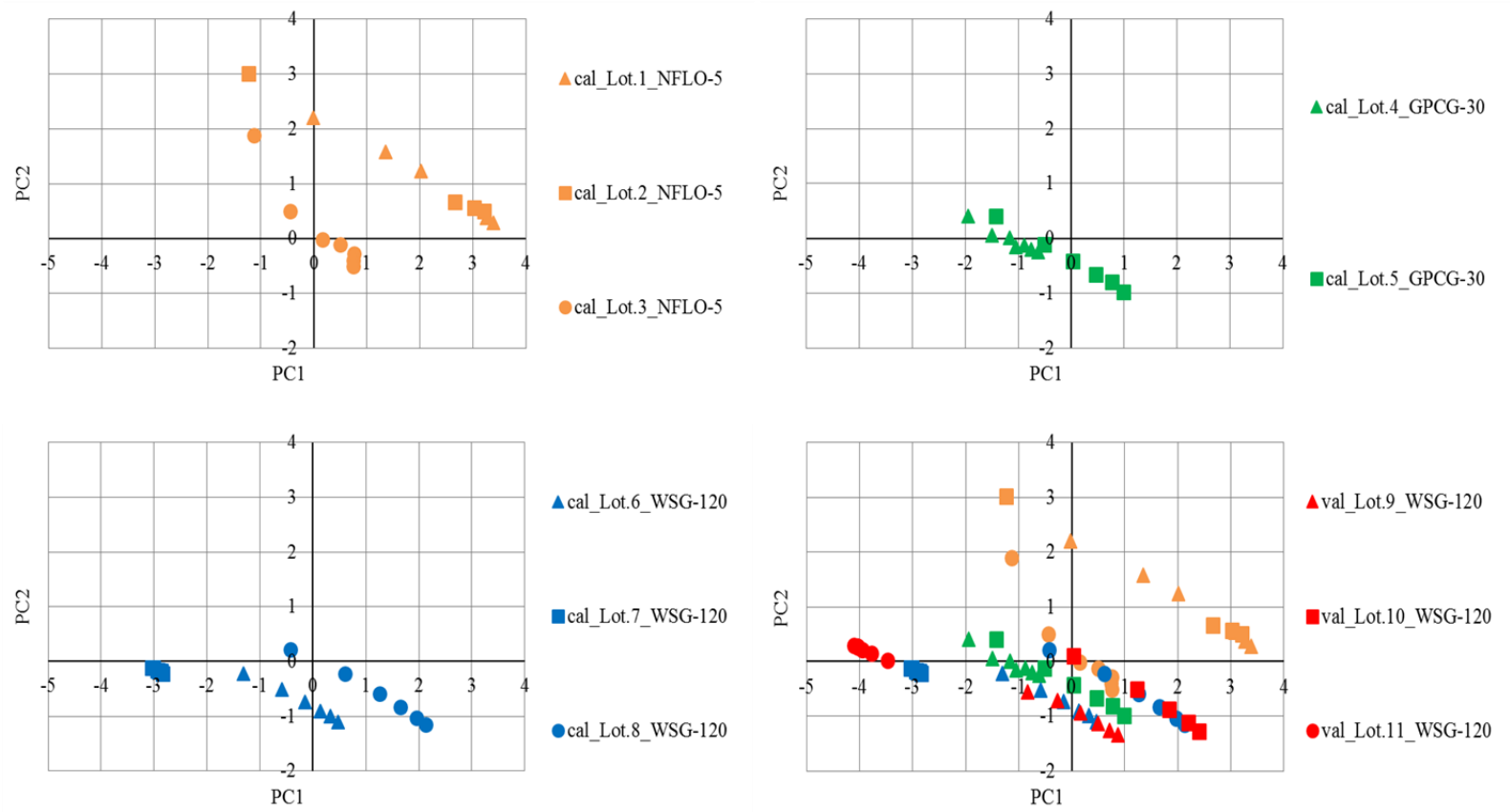


Figure 4.3. Scatter plots with the first and second PC in PCA conducted using calibration dataset H, which was utilized to construct a PP-based LW-PLSR model.

Orange, green, and blue symbols are calibration samples acquired at laboratory, pilot, and commercial scale, respectively. Red symbol indicates validation samples projected onto the subspace spanned by the first and second PC.

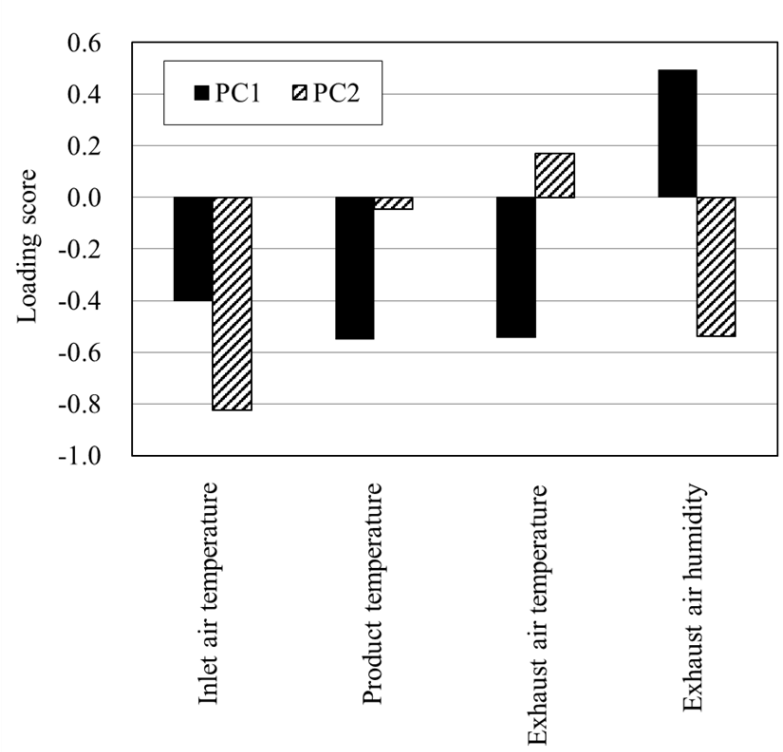


Figure 4.4. Loading scores of individual PPs in PCA conducted using calibration dataset H, which was utilized to construct a PP-based LW-PLSR model.

Besides, Figure 4.1 shows that the $RMSE_P$ of 2.0% for the model constructed by dataset I was larger than that of 0.5% for the model built using dataset E. As shown in Table 4.1, dataset I consisted of two pilot-scale lots, while dataset E contained three laboratory-scale lots and two pilot-scale lots. Although the calibration samples, Lot Nos.1 and 2, were far from the validation samples compared to the other calibration samples (Lot Nos. 3, 4, and 5), they contributed to improving the prediction accuracy. Thus, it was concluded that the construction of accurate PP-based LW-PLSR models required the calibration samples with the following two features: 1) located near the validation samples on the subspace spanned by PCs, and 2) having a wide range of variations in PC scores. As provided in Figure 4.4, the inlet air temperature and exhaust air humidity

affected both the first and second PCs, while the product temperature and exhaust air temperature mainly had an impact on the first PC. The product temperature, exhaust air temperature, and exhaust air humidity correlate with the granule water content regardless of the manufacturing scale, as shown in Figure 4.5. These results indicate that the distance between samples on the PC1-PC2 subspace derives from the difference of inlet air temperature and granule water content. The inlet air temperature is an operable PP, and the granule water content is controlled by other two operable PPs, *i.e.*, inlet air volume and spray rate. Hence, regardless of the manufacturing scale, it is possible to obtain the calibration samples located near the validation samples on the PC1-PC2 subspace by setting the inlet air temperature and granule water content to align with the validation samples. Also, we can acquire different samples on the PC1-PC2 subspace by changing inlet air temperature and granule water content independently, which contributes to the preparation of the calibration samples having a wide range of variations in PC scores.

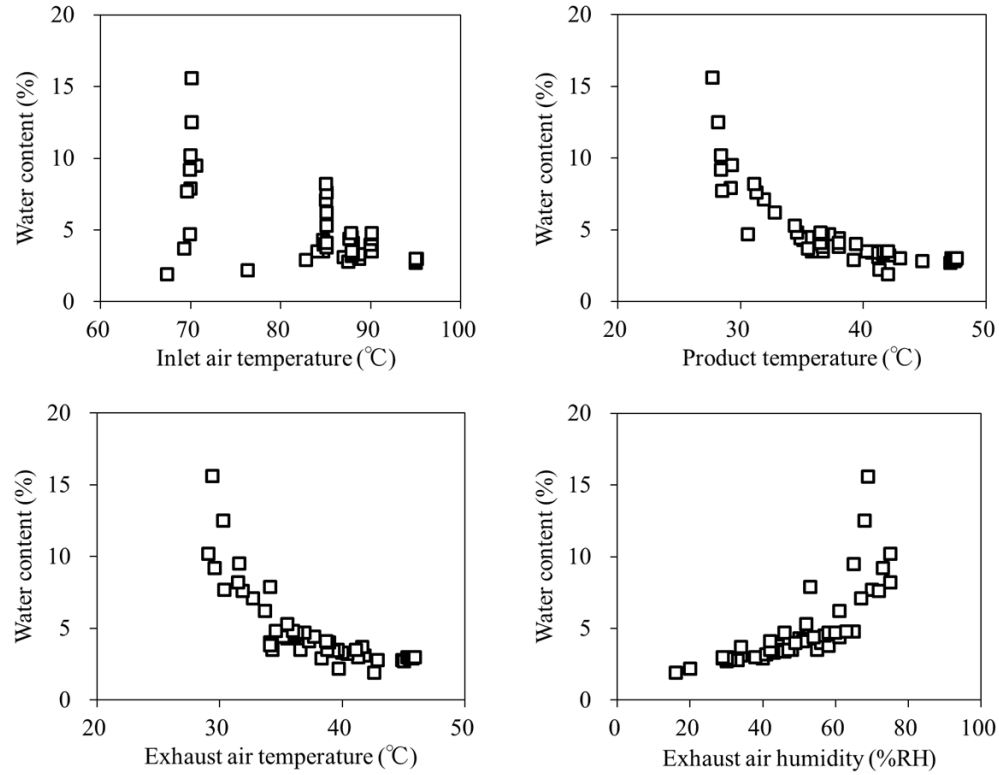


Figure 4.5. Relationship between each PP and water content of granules in calibration dataset H, which was utilized to construct a PP-based LW-PLSR model.

### 4.3.1.2. NIRS-based LW-PLSR models

Figure 4.6 shows the $RMSE_P$ values of the NIRS-based LW-PLSR models constructed utilizing the absorbance at the selected wavenumber region of the NIR spectra, *i.e.*, from 7212 $cm^{-1}$ to 6935 $cm^{-1}$. As shown in Figure 4.6, the LW-PLSR model constructed using three laboratory-scale lots, *i.e.*, calibration dataset C, fulfilled the criterion of $RMSE_P$. Hence, the values of $N_{lab}$, $N_{pilot}$, and $N_{coml}$ were determined to be three, zero, and zero, respectively.

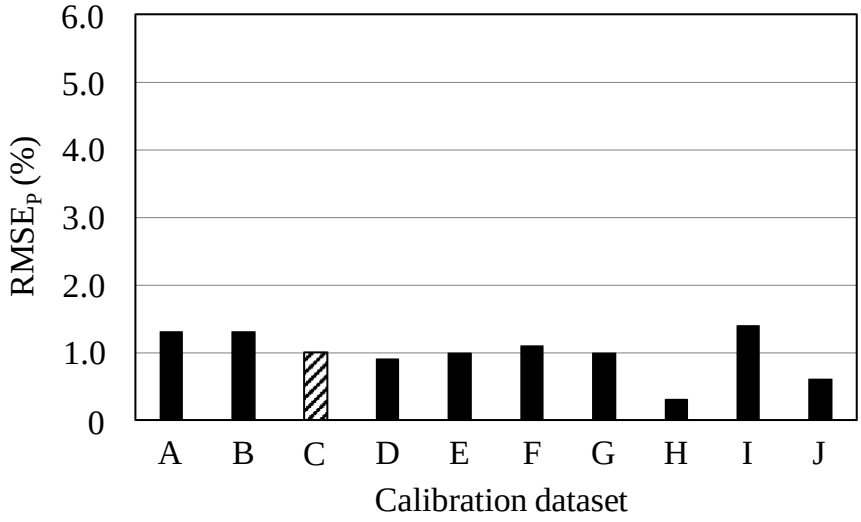


Figure 4.6. Prediction accuracy of NIRS-based LW-PLSR models.

The diagonal bar indicates the calibration dataset with the minimum required number of experimental data to meet the criterion of $RMSE_P$.

PCA was applied to the calibration dataset H. The relationship between the number of PCs and the cumulative proportion is shown in Figure 4.2. With the first PC, the cumulative proportion surpassed 0.9. The scatter plots of the first and second PCs are provided in Figure 4.7. Most of the calibration samples existed near the validation samples, which resulted in the $RMSE_P$ values of all the models to be around 1.0. It was suggested that the construction of accurate NIRS-based LW-PLSR models requires the calibration samples with the same features as PP-based LW-PLSR models.

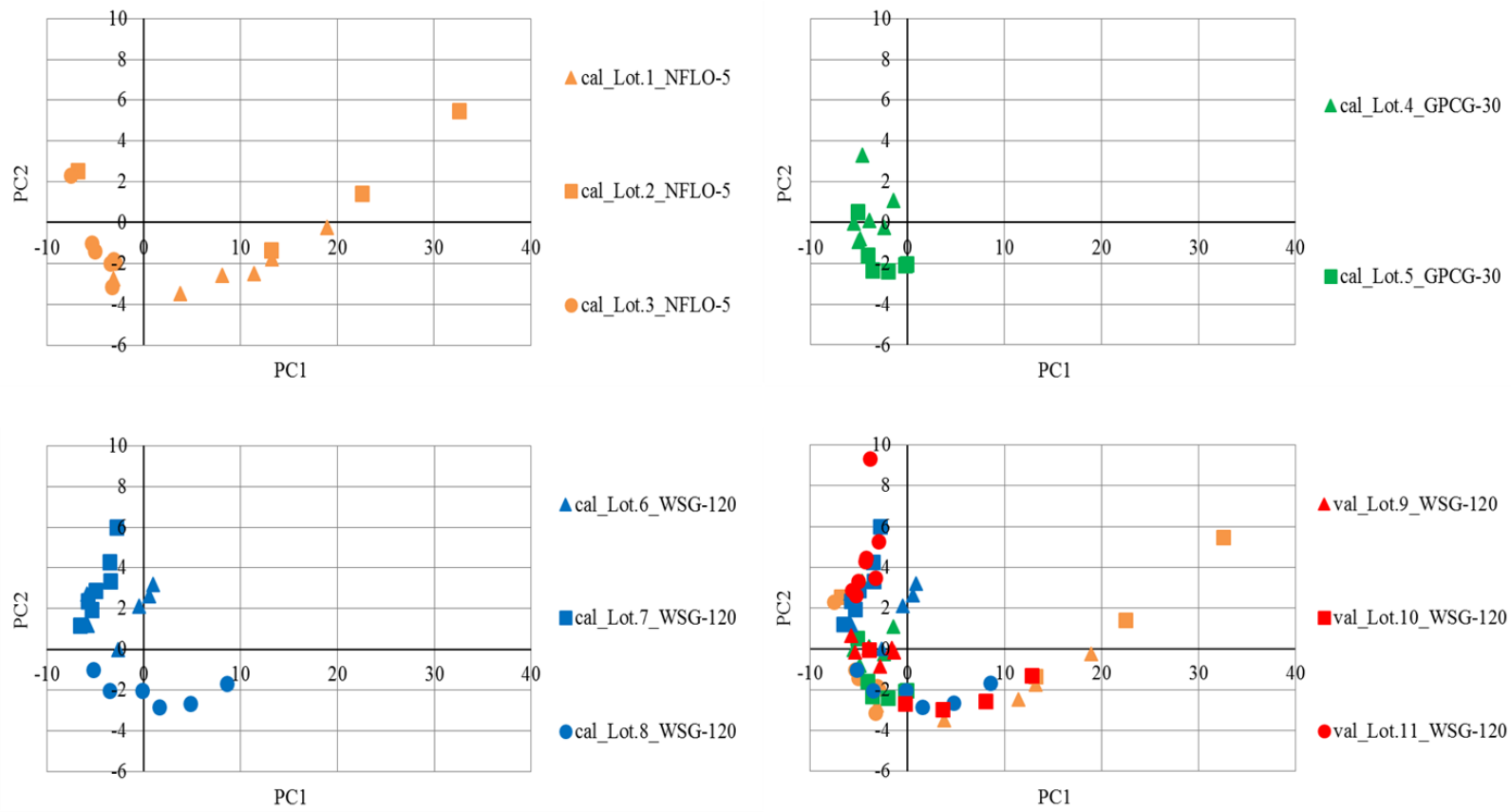


Figure 4.7. Scatter plots with the first and second PC in PCA conducted using calibration dataset H, which was utilized to construct a NIRS-based LW-PLSR model.

Orange, green, and blue symbols are calibration samples acquired at laboratory, pilot, and commercial scale, respectively. Red symbol indicates validation samples projected onto the subspace spanned by the first and second PC.

Figure 4.8 indicates that PC1, which was the dominant PC with a high proportion of 0.91, strongly correlated with the water content of granules, regardless of the manufacturing scale. This result means that the NIR spectra of the granules are dependent on the water content, regardless of the manufacturing equipment. Thus, it is possible to obtain calibration samples having desired scores of the dominant PC by setting the calibrated range of granule water content to align with the range of the validation samples.

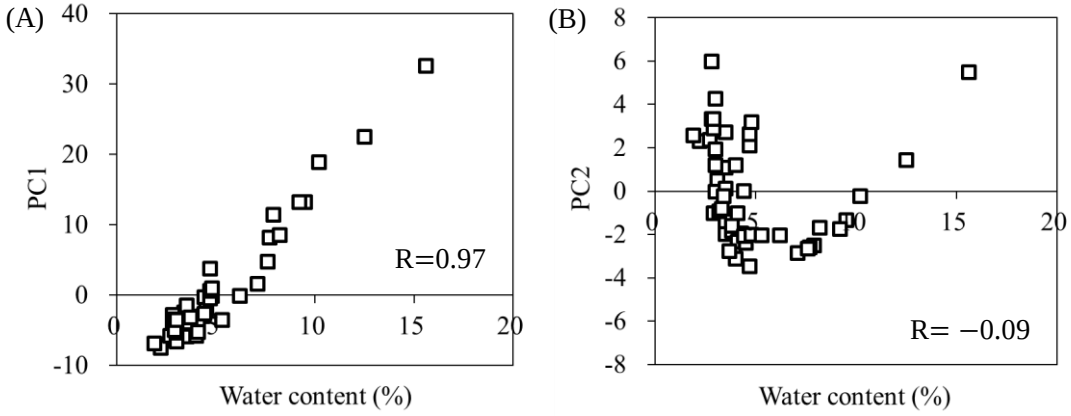


Figure 4.8. Influence of water content of granules on the first (A) and second (B) PC in PCA conducted using calibration dataset H, which was utilized to construct a NIRS-based LW-PLSR model.

## 4.3.2. Comparison of development cost of PAT

$Cost_{rm}$ of PP-based and NIRS-based LW-PLSR models were calculated through Eqs. (4.2), (4.3), (4.4), and (4.5) at the determined values of $N_{lab}$, $N_{pilot}$, and $N_{coml}$ (refer to sections 4.3.1.1. and 4.3.1.2.). $Cost_{dev}$ of both models were computed using $Cost_{rm}$ and the preset values of $Cost_{inv}$ according to Eq. (4.1). Their difference $Cost_{difference}$ was calculated by Eq. (4.6). Figure 4.9 shows the value of $Cost_{difference}$ depending on

$Cost_{API}$, $MF_{API}$, and $Cost_{inv}$. As shown in Figure 4.9, the reduction in $Cost_{API}$ and $MF_{API}$ made the PP-based LW-PLSR models more cost-effective compared with the NIRS-based LW-PLSR models. This is because the impact of $Cost_{rm}$ on $Cost_{dev}$ became smaller than that of $Cost_{inv}$ on $Cost_{dev}$ (refer to Eq. (4.1)). Besides, the area where PP-based LW-PLSR models become more cost-effective was enlarged, along with the increase in $Cost_{inv}$.

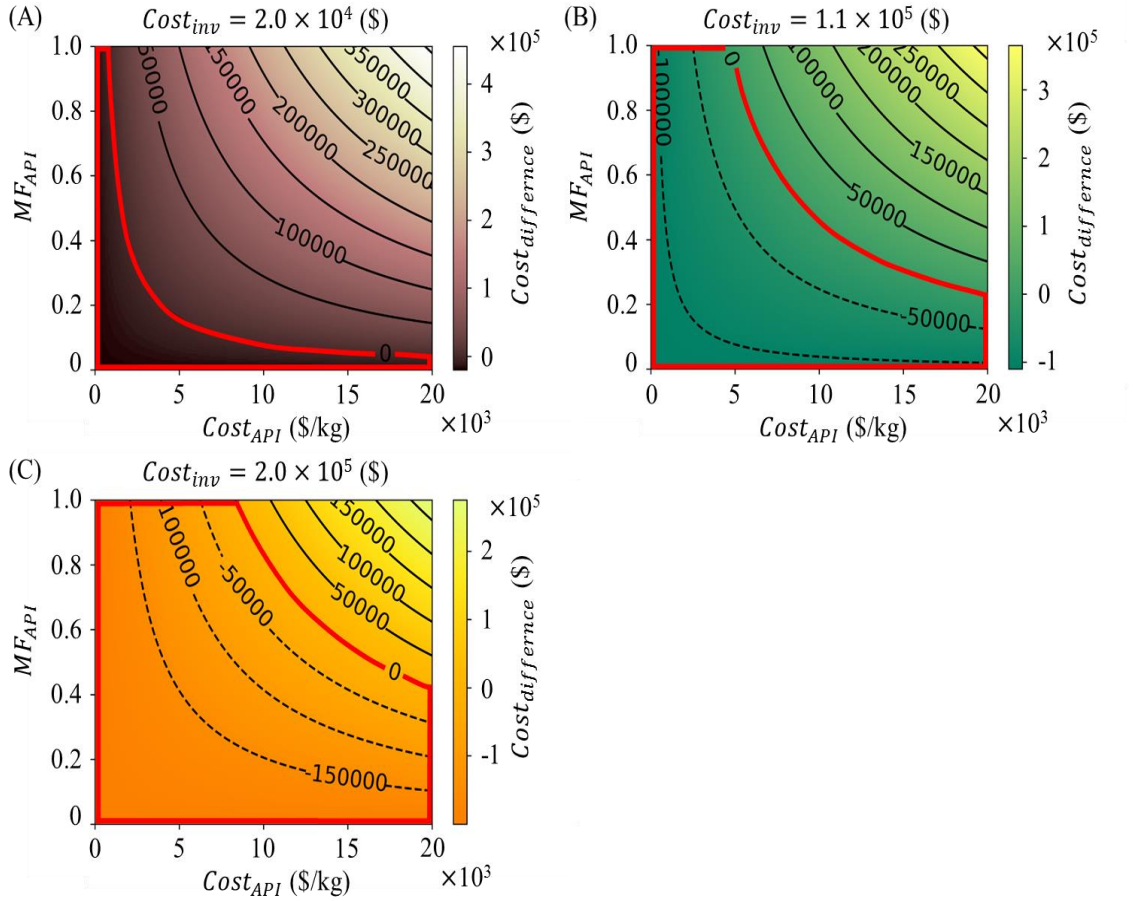


Figure 4.9. $Cost_{difference}$ depending on $Cost_{API}$, $MF_{API}$, and $Cost_{inv}$.

$Cost_{inv}$, the price of the NIR spectrometer, was set as (A) \$2.0×10⁴, (B) \$1.1×10⁵, and (C) \$2.0×10⁵. The area surrounded by red line indicates $Cost_{dev,PP}$ is smaller than $Cost_{dev,NIRS}$.

## 4.4. Summary

In Chapter 4, two types of statistical models to predict water content, *i.e.*, PP-based and NIRS-based LW-PLSR models, were evaluated with respect to prediction accuracy and development cost. This study demonstrated that both PP-based and NIRS-based LW-PLSR models achieved high prediction accuracy on the commercial scale, without commercial-scale experimental data in the calibration dataset.

This study clarified the key points to construct accurate PP-based and NIRS-based LW-PLSR models. It is crucial to prepare the calibration samples with the following two features: 1) located near the validation samples on the subspace spanned by PCs, and 2) having a wide range of variations in PC scores. In PP-based LW-PLSR models, scores of the dominant PCs depend on inlet air temperature and granule water content regardless of the manufacturing scale. Thus, it is possible to acquire calibration samples with desired PC scores by adjusting inlet air temperature and granule water content independently. In contrast, NIRS-based LW-PLSR models are simple; scores of the dominant PC strongly correlate with granule water content regardless of the manufacturing scale.

Besides, this study revealed which type of LW-PLSR model should be selected depending on the raw material cost, composition of granules, and price of the NIR spectrometer. The reduction in cost and mass fraction of API makes the PP-based LW-PLSR models more cost-effective than the NIRS-based LW-PLSR models. Additionally, the PP-based LW-PLSR models become more cost-effective as the price of the NIR spectrometer increases.

The revealed points through this study are expected to promote the efficiency of building accurate models and save the development cost of PAT.

# 5. Gray-box Soft Sensor

## 5.1. Introduction

This study aims to develop descriptive and accurate gray-box models for water content monitoring in fluidized bed granulation. These gray-box models enable to promote process understanding and more accurate quality control. In this study, three types of gray-box models, *i.e.*, parallel,[52] serial,[53] and combined[44] gray-box models (see Figure 1.2), were constructed by integrating the heat and mass balance model (white-box model) and the LW-PLSR model (black-box model). This study evaluated the prediction accuracy of the white-box, parallel gray-box, serial gray-box, and combined gray-box models, which were built using real operating data on a commercial scale with two formulations.

## 5.2. Materials and Methods

### 5.2.1. Materials

The two formulations A and B were granulated using the commercial-scale fluidized bed granulator, refer to Sections 2.1 and 2.2.

### 5.2.2. Measurement of PPs and reference

Refer to Sections 2.3 and 2.6.

### 5.2.3. Calibration and validation datasets

As shown in Table 2.2, the six granulation lot data were prepared in each formulation. Four granulation lots were selected from the six lots as a calibration dataset. The remaining two granulation lots were utilized as a validation dataset. Since lot selection impacts the prediction accuracy, this study evaluated all 15 combinations of the calibration and validation datasets. This procedure was common to both formulations, as provided in Table 5.1.

Table 5.1. Combination of calibration and validation datasets.

| Trial No. | Lot No. | |
|---|---|---|
| | Calibration dataset | Validation dataset |
| 1 | 3, 4, 5, 6 | 1, 2 |
| 2 | 2, 4, 5, 6 | 1, 3 |
| 3 | 2, 3, 5, 6 | 1, 4 |
| 4 | 2, 3, 4, 6 | 1, 5 |
| 5 | 2, 3, 4, 5 | 1, 6 |
| 6 | 1, 4, 5, 6 | 2, 3 |
| 7 | 1, 3, 5, 6 | 2, 4 |
| 8 | 1, 3, 4, 6 | 2, 5 |
| 9 | 1, 3, 4, 5 | 2, 6 |
| 10 | 1, 2, 5, 6 | 3, 4 |
| 11 | 1, 2, 4, 6 | 3, 5 |
| 12 | 1, 2, 4, 5 | 3, 6 |
| 13 | 1, 2, 3, 6 | 4, 5 |
| 14 | 1, 2, 3, 5 | 4, 6 |
| 15 | 1, 2, 3, 4 | 5, 6 |

### 5.2.4. Heat and mass balance model (white-box model)

The present work adopted the main principles of the heat and mass balance model developed by Tanino *et al.*[10] However, this study modified several equations and introduced three fitting parameters, *i.e.*, $\beta$, $\gamma$, and $\delta$, to describe the phenomena during granulation more realistically.

### 5.2.4.1. Equations for water content calculation

The granule water content of wet granules at $t$ min (%) was calculated using the following equation:

$$M(t) = \frac{W(t)}{S(t)} \times 100 \tag{5.1}$$

where $W(t)$ is water content quantity in wet granules (kg), and $S(t)$ is total weight of wet granules (kg). $W(t)$ and $S(t)$ are expressed as follows:

$$W(t) = \int \left( AL_w(t) - G_w(t) \right) dt + W(0) \tag{5.2}$$

$$S(t) = \int \left( L_w(t) - G_w(t) \right) dt + S(0) \tag{5.3}$$

where $A$ is water ratio of binder solution, $L_w(t)$ is spray rate (kg/min), and $G_w(t)$ is water evaporation speed (kg/min). The setting values of $A$, $W(0)$, and $S(0)$ in formulations A and B are listed in Table 5.2. $G_w(t)$ is expressed as follows:

$$G_w(t) = \frac{\alpha}{H(t)} \left( E(t) - E_{warm}(t) \right) \tag{5.4}$$

$$H(t) = 539 + 0.56 \left( 100 - T_p(t) \right) \tag{5.5}$$

$$E(t) = \left( T_f(t) - T_e(t) + \beta \right) \gamma V_f(t) \rho(t) k_a \tag{5.6}$$

$$E_{warm}(t) = \delta L_w(t) k_w \tag{5.7}$$

$$\rho(t) = \frac{273}{273 + T_f(t)} \times \frac{29}{22.4} \tag{5.8}$$

where $\alpha$ is drying efficiency, which is a fitting parameter ranging from 0 to 1, $H(t)$ is heat of water vaporization (kcal/kg), $T_p(t)$ is product temperature (℃), $E(t)$ is the difference in heat quantity between inlet and exhaust air (kcal/min), and $E_{warm}(t)$ is sensible heat utilized to warm the water in the binder solution (kcal/min), $T_f(t)$ is inlet

air temperature (℃), $T_e(t)$ is exhaust air temperature (℃), $V_f(t)$ is inlet air volume ($m^3$/min), and $\rho(t)$ is air density (kg/$m^3$). $k_a$ is specific heat of air (kcal/kg· ℃), and $k_w$ is specific heat of water (kcal/kg· ℃). In this study, $k_a$ and $k_w$ were set to 0.238 kcal/kg· ℃ and 1 kcal/kg· ℃, respectively.[66)]

For the calculation of Eq. (5.6), $T_f(t)$, $T_e(t)$, and $V_f(t)$ should be assigned the values just before flowing in or after flowing out through the granulator (denoted as "ideal values"). However, the measured values could be different from the ideal values due to the positions of the sensors in the duct. For example, the inlet air temperature measured by the thermometer located far from the air supply opening could be higher than the ideal value. To adjust the gap between the measured and ideal values, this study introduced two fitting parameters of $\beta$ and $\gamma$; $\beta$ ranged from -25 ℃ to 0 ℃, and $\gamma$ ranged from 0.5 to 1. In addition, $\delta$ is the temperature difference between granule surface at water evaporation and binder solution, which is a fitting parameter ranging from 0 ℃ to 25 ℃.

Table 5.2. Setting values of $A$, $W(0)$, and $S(0)$ in the heat and mass balance model.

| Formulation A | | | Formulation B | | |
|---|---|---|---|---|---|
| $A$ | $W(0)$ (kg) | $S(0)$ (kg) | $A$ | $W(0)$ (kg) | $S(0)$ (kg) |
| 0.991 | 1.75 | 97.3 | 0.930 | 2.41 | 96.2 |

## 5.2.4.2. Fitting parameters optimization

The four fitting parameters in the heat and mass balance model, *i.e.*, $\alpha$, $\beta$, $\gamma$, and $\delta$, were optimized through the following steps:

I. Construct a heat and mass balance model $f_{\mathrm{hm}}$ to predict an output variable $y$.

$$\hat{y}_{\mathrm{hm},n} = f_{\mathrm{hm}}\left(\boldsymbol{x}_{\mathrm{hm},n}, \boldsymbol{\theta}\right) \tag{5.9}$$

where $\hat{y}_{\mathrm{hm},n}$ is the granule water content predicted by $f_{\mathrm{hm}}$ for the $n$-th sample. $\boldsymbol{x}_{\mathrm{hm},n}$ is the corresponding inputs: spraying time (min), mean value of spray rate for 0 min to $t$ min (kg/min), product temperature (°C), inlet air temperature (°C), exhaust air temperature (°C), and inlet air volume ($m^3$/min). $\boldsymbol{\theta}$ is a fitting parameters vector, whose elements are $\alpha$, $\beta$, $\gamma$, and $\delta$.

II. Determine $\boldsymbol{\theta}$ to minimize the prediction error sums of squares using Bayesian optimization.

$$\tilde{\boldsymbol{\theta}} = \arg\min_{\boldsymbol{\theta}} \sum_{n=1}^{N} e_{\mathrm{hm},n}^2 \tag{5.10}$$

$$e_{\mathrm{hm},n} = y_n - f_{\mathrm{hm}}\left(\boldsymbol{x}_{\mathrm{hm,n}}, \boldsymbol{\theta}\right) \tag{5.11}$$

$$\boldsymbol{\theta}_{\mathrm{L}} \le \boldsymbol{\theta} \le \boldsymbol{\theta}_{\mathrm{U}} \tag{5.12}$$

where $e_{\mathrm{hm},n}$ is the output error of the heat and mass balance model, and $y_n$ is the output measurement for the $n$-th sample. $N$ is the number of samples in the calibration dataset. $\boldsymbol{\theta}_{\mathrm{L}}$ and $\boldsymbol{\theta}_{\mathrm{U}}$ are lower and upper limit vectors, respectively. The ranges of $\alpha$, $\beta$, $\gamma$, and $\delta$ are provided in Section 5.2.4.1. Calculations in Bayesian optimization were conducted using the Statistics and Machine Learning Toolbox in MATLAB® software (MathWorks, Inc.), and the acquisition function 'expected-improvement-plus' was selected from the options.

### 5.2.5. Gray-box models

In this section, the modeling methods of parallel, serial, and combined gray-box models are described. The general framework of gray-box modeling and the fundamental concept of parallel, serial, and combined gray-box models have been previously reported elsewhere.[44,45)]

#### 5.2.5.1. Parallel gray-box model

Parallel gray-box models utilize the black-box model to offset the output error of the white-box model. The parallel gray-box models were constructed through the following steps:

I. Build a heat and mass balance model $\hat{y}_{\mathrm{hm},n} = f_{\mathrm{hm}}\left(\boldsymbol{x}_{\mathrm{hm},n}, \boldsymbol{\theta}\right)$.

II. Optimize fitting parameters $\boldsymbol{\theta}$ through Eqs. (5.10), (5.11), and (5.12).

III. Construct a statistical model $f_{\mathrm{pa}}$ to estimate the output error of the heat and mass balance model $e_{\mathrm{hm},n}$.

$$\tilde{\boldsymbol{\varphi}}_{\mathrm{pa}} = \arg\min_{\boldsymbol{\varphi}_{\mathrm{pa}}} \sum_{n=1}^{N} \left(e_{\mathrm{hm},n} - f_{\mathrm{pa}}\left(\boldsymbol{x}_{\mathrm{sm},n}, \boldsymbol{\varphi}_{\mathrm{pa}}\right)\right)^2 \tag{5.13}$$

$$\hat{e}_{\mathrm{hm},n} = f_{\mathrm{pa}}\left(\boldsymbol{x}_{\mathrm{sm},n}, \boldsymbol{\varphi}_{\mathrm{pa}}\right) \tag{5.14}$$

where $\boldsymbol{\varphi}_{\mathrm{pa}}$ is a vector of tuning parameters in the LW-PLSR model $f_{\mathrm{pa}}$, *i.e.*, $R$ (the number of latent variables) and $\eta$ (the localization parameter), refer to Section 2.7.1. The localization parameter $\eta$ was optimized in the range of 0 to 10.0. When the localization parameter $\eta$ is 0, LW-PLSR is equivalent to PLSR. $\boldsymbol{x}_{\mathrm{sm},n}$ is the $n$-th measurement vector of nine inputs: spraying time (min), inlet air volume ($m^3$/min),

inlet air temperature (°C), inlet air humidity (g-water/kg-air), mean value of spray rate for 0 min to $t$ min (kg/min), spray air volume (NL/min), product temperature (°C), exhaust air temperature (°C), and exhaust air humidity (%RH).

IV. Develop a gray-box model by integrating the heat and mass balance model and the statistical model $f_{\mathrm{pa}}$.

$$\hat{y}_{\mathrm{pa},n} = f_{\mathrm{hm}}\left(\boldsymbol{x}_{\mathrm{hm},n}, \tilde{\boldsymbol{\theta}}\right) + f_{\mathrm{pa}}\left(\boldsymbol{x}_{\mathrm{sm},n}, \tilde{\boldsymbol{\varphi}}_{\mathrm{pa}}\right) \tag{5.15}$$

where $\hat{y}_{\mathrm{pa},n}$ is the $n$-th prediction value obtained using the parallel gray-box model.

### 5.2.5.2. Serial gray-box model

Serial gray-box models use the black-box model to update the fitting parameters of the white-box model based on the PPs. The LW-PLSR model predicted all four fitting parameters of the heat and mass balance model, *i.e.*, $\alpha$, $\beta$, $\gamma$, and $\delta$, because they depend on manufacturing conditions. The serial gray-box models were built through the following steps:

I. Build a heat and mass balance model $\hat{y}_{\mathrm{hm},n} = f_{\mathrm{hm}}\left(\boldsymbol{x}_{\mathrm{hm},n}, \boldsymbol{\theta}\right)$.

II. Determine fitting parameters $\boldsymbol{\theta}$ for each calibration sample using Bayesian optimization.

$$\tilde{\boldsymbol{\theta}}_n = \arg\min_{\boldsymbol{\theta}_n}\left(y_n - f_{\mathrm{hm}}\left(\boldsymbol{x}_{\mathrm{hm},n}, \boldsymbol{\theta}_n\right)\right)^2 \qquad \left(n = 1, 2, 3, ..., N\right) \tag{5.16}$$

$$\boldsymbol{\theta}_{\mathrm{L}} \le \boldsymbol{\theta}_n \le \boldsymbol{\theta}_{\mathrm{U}} \tag{5.17}$$

where $\tilde{\boldsymbol{\theta}}_n$ is the optimal fitting parameters vector for the $n$-th sample.

III. Construct a statistical model $f_{\mathrm{se}}$ to predict $\tilde{\boldsymbol{\theta}}_n$. When $\hat{\tilde{\boldsymbol{\theta}}}_n$ was lower than $\boldsymbol{\theta}_{\mathrm{L}}$ or higher than $\boldsymbol{\theta}_{\mathrm{U}}$, $\hat{\tilde{\boldsymbol{\theta}}}_n$ was replaced with $\boldsymbol{\theta}_{\mathrm{L}}$ or $\boldsymbol{\theta}_{\mathrm{U}}$, respectively.

$$\tilde{\boldsymbol{\varphi}}_{\mathrm{se}} = \arg\min_{\boldsymbol{\varphi}_{\mathrm{se}}} \sum_{n=1}^{N} \left( \tilde{\boldsymbol{\theta}}_n - f_{\mathrm{se}} \left( \boldsymbol{x}_{\mathrm{sm},n}, \boldsymbol{\varphi}_{\mathrm{se}} \right) \right)^2 \tag{5.18}$$

$$\hat{\tilde{\boldsymbol{\theta}}}_n = f_{\mathrm{se}} \left( \boldsymbol{x}_{\mathrm{sm},n}, \boldsymbol{\varphi}_{\mathrm{se}} \right) \tag{5.19}$$

where $\boldsymbol{\varphi}_{\mathrm{se}}$ is a vector of tuning parameters in the LW-PLSR model $f_{\mathrm{se}}$.

IV. Develop a gray-box model by integrating the heat and mass balance model and the statistical model $f_{\mathrm{se}}$.

$$\hat{y}_{\mathrm{se},n} = f_{\mathrm{hm}} \left( \boldsymbol{x}_{\mathrm{hm},n}, f_{\mathrm{se}} \left( \boldsymbol{x}_{\mathrm{sm},n}, \tilde{\boldsymbol{\varphi}}_{\mathrm{se}} \right) \right) \tag{5.20}$$

where $\hat{y}_{\mathrm{se},n}$ is the $n$-th prediction value based on the serial gray-box model.

### 5.2.5.3. Combined gray-box model

Combined gray-box models utilize both a serial gray-box model and a parallel gray-box model; in other words, the black-box model is used to compensate for the output error of the serial gray-box model. The combined gray-box models were developed by applying a part of the model-building procedure of the parallel gray-box model, *i.e.*, steps III and IV, to the serial gray-box model.

### 5.2.6. Prediction performance evaluation

In the white-box models, the $RMSE_C$ of the granule water content was calculated. In the parallel, serial, and combined gray-box models, the $RMSE_{CV}$ was calculated.

To evaluate the applicability to real operating data on a commercial scale, the $RMSE_P$ and $R^2$ were calculated using the validation dataset, which was not included in the calibration dataset. The $RMSE_P$ criterion required for application to commercial production was set to 1.0%, based on the study in Chapter 4. The detailed information on calculation is described in Section 2.8.

### 5.2.7. Evaluation of LW-PLSR model

In Chapter 3, it was demonstrated that the assessment based on $T^2$ and Q was also valuable for testing whether the LW-PLSR model was valid for the query. A 99% confidence limit was adopted as the threshold of $T^2$ and Q to align with the study in Chapter 3. The detailed information on calculation is described in Section 2.9.

## 5.3. Results and discussion

### 5.3.1. Model building

In both formulations, the white-box model and three types of gray-box models were constructed 15 times with different calibration datasets. Their prediction accuracy was evaluated based on the median values of the RMSE. In this section, the parameter optimization and LOOCV results of each model are described.

#### 5.3.1.1. Heat and mass balance model (white-box model)

The fitting parameter values determined by Bayesian optimization considering all the calibration samples are shown in Figure 5.1. Regardless of the formulation, the optimal values of $\delta$ varied remarkably, depending on Trial No. This result is reasonable because the temperature of the granule surface at water evaporation depends on the manufacturing conditions. Figure 5.2 shows that the median values of $RMSE_C$ in formulations A and B were 1.6% and 1.0%, respectively. The accuracy did not satisfy the requirement.

#### 5.3.1.2. Parallel gray-box model

The tuning parameters of the LW-PLSR models determined based on LOOCV are listed in Table 5.3. As shown in Figure 5.2, the median values of $RMSE_{CV}$ for the parallel gray-box models in formulations A and B were 0.6% and 0.2%, respectively. They were 63% and 80% smaller than those of $RMSE_C$ in the white-box models, *i.e.*, 1.6% and 1.0%, respectively. This result means that the LW-PLSR models successfully compensated for the output error of the white-box models, and the parallel gray-box modeling was useful for improving the prediction accuracy.

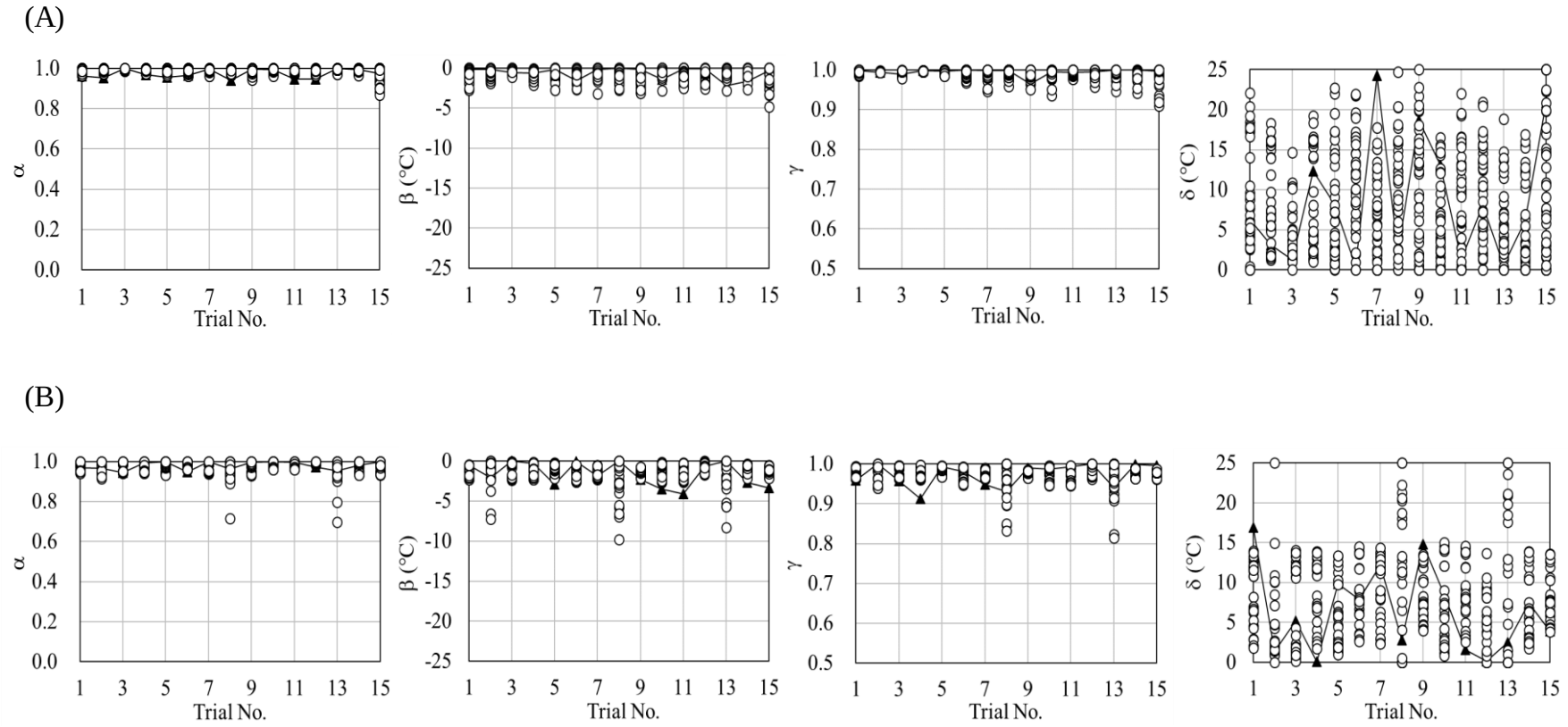


Figure 5.1. Fitting parameter values of the heat and mass balance model, *i.e.*, $\alpha$, $\beta$, $\gamma$, and $\delta$, in case of (A) formulation A and (B) formulation B.

In white-box models (filled triangles), the fitting parameter values were determined by Bayesian optimization considering all calibration samples. In serial gray-box models (open circles), the fitting parameter value for each query was predicted using the LW-PLSR model in LOOCV.

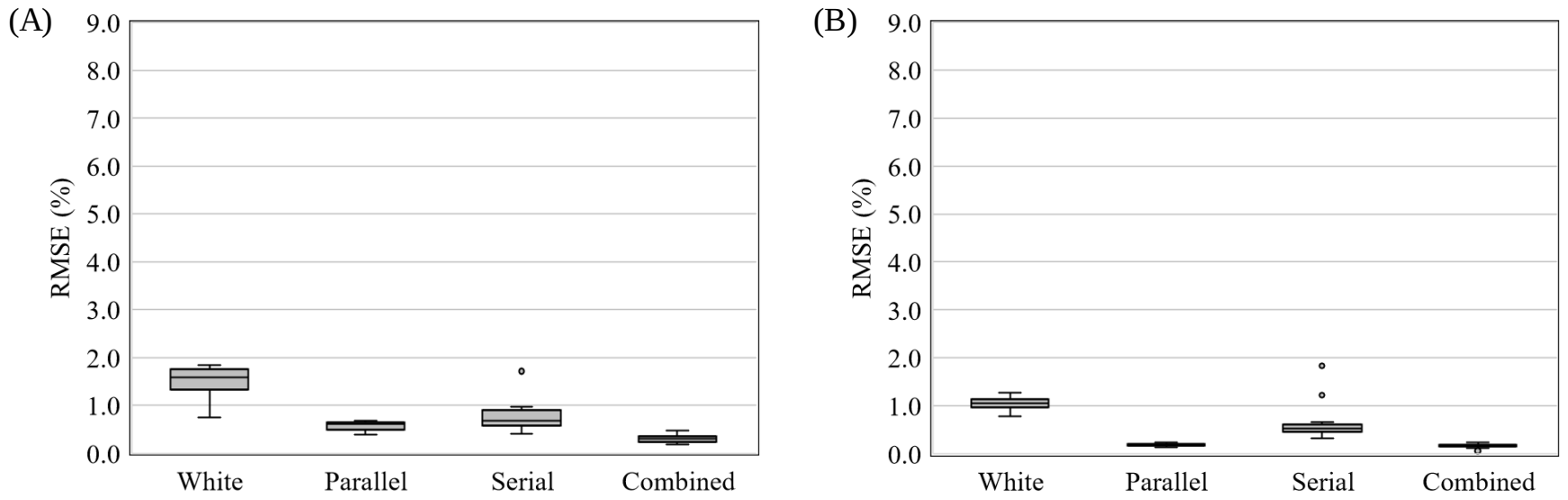


Figure 5.2. RMSE boxplots of white-box, parallel gray-box, serial gray-box, and combined gray-box models in case of (A) formulation A and (B) formulation B.

In white-box models, $RMSE_C$ was described. In gray-box models, LOOCV was conducted, and $RMSE_{CV}$ was calculated.

Table 5.3. Tuning parameters of LW-PLSR models determined by LOOCV.

| Formulation | Trial No. | Parallel gray-box model | | Serial gray-box model | | Combined gray-box model | |
|---|---|---|---|---|---|---|---|
| | | Number of latent variables | Localization parameter | Number of latent variables | Localization parameter | Number of latent variables | Localization parameter |
| A | 1 | 7 | 2.0 | 6 | 0.2 | 4 | 2.3 |
| A | 2 | 8 | 1.5 | 1 | 0 | 5 | 1.2 |
| A | 3 | 4 | 1.2 | 1 | 1.5 | 2 | 0.6 |
| A | 4 | 7 | 1.9 | 1 | 0 | 4 | 2.5 |
| A | 5 | 9 | 2.3 | 4 | 0.5 | 9 | 3.2 |
| A | 6 | 5 | 1.2 | 6 | 0 | 4 | 0.1 |
| A | 7 | 6 | 5.5 | 6 | 1.1 | 5 | 1.1 |
| A | 8 | 5 | 2.1 | 5 | 0.2 | 5 | 2.4 |
| A | 9 | 6 | 2.1 | 9 | 1.1 | 6 | 2.4 |
| A | 10 | 8 | 0.4 | 3 | 1.3 | 8 | 0 |
| A | 11 | 5 | 1.6 | 4 | 0 | 5 | 2.3 |
| A | 12 | 5 | 1.5 | 2 | 0.9 | 6 | 1.4 |
| A | 13 | 7 | 1.4 | 7 | 0.5 | 4 | 1.2 |
| A | 14 | 6 | 1.0 | 4 | 1.3 | 9 | 1.5 |
| A | 15 | 5 | 2.2 | 8 | 9.0 | 1 | 1.3 |
| B | 1 | 4 | 6.5 | 1 | 0 | 4 | 9.8 |
| B | 2 | 2 | 2.4 | 2 | 10.0 | 9 | 1.6 |
| B | 3 | 4 | 3.8 | 1 | 0.2 | 4 | 3.8 |
| B | 4 | 4 | 7.5 | 1 | 0 | 4 | 5.5 |
| B | 5 | 5 | 4.5 | 1 | 0 | 4 | 4.2 |
| B | 6 | 3 | 4.6 | 1 | 0.1 | 2 | 5.5 |
| B | 7 | 3 | 4.7 | 1 | 0.5 | 3 | 3.2 |
| B | 8 | 9 | 2.8 | 9 | 5.8 | 1 | 0 |
| B | 9 | 5 | 5.3 | 1 | 0 | 5 | 4.9 |
| B | 10 | 3 | 4.1 | 1 | 0 | 2 | 4.2 |
| B | 11 | 3 | 2.9 | 1 | 0 | 3 | 2.8 |
| B | 12 | 8 | 2.8 | 2 | 0 | 7 | 2.3 |
| B | 13 | 5 | 1.5 | 9 | 6.2 | 2 | 2.3 |
| B | 14 | 4 | 2.0 | 1 | 0 | 3 | 1.6 |
| B | 15 | 4 | 2.1 | 1 | 0 | 3 | 2.1 |

### 5.3.1.3. Serial gray-box model

Figure 5.1 shows the fitting parameter value for each query predicted using the LW-PLSR model in LOOCV. The predicted value of $\delta$ differed significantly, depending on the manufacturing condition of each query. Table 5.3 shows the tuning parameters of the LW-PLSR models determined based on LOOCV. As shown in Figure 5.2, the median values of $RMSE_{CV}$ for the serial gray-box models in formulations A and B were 0.7% and 0.5%, respectively. They were 56% and 50% smaller than those of $RMSE_C$ in the white-box models, *i.e.*, 1.6% and 1.0%, respectively. This result indicates that the LW-PLSR models adjusted the fitting parameters of the white-box model depending on the PPs, which made the serial gray-box models more accurate than the white-box models.

### 5.3.1.4. Combined gray-box model

Table 5.3 lists the tuning parameters of the LW-PLSR models determined based on LOOCV. Figure 5.2 shows that the median values of $RMSE_{CV}$ in formulations A and B were 0.3% and 0.2%, respectively. They were 57% and 60% smaller than those of $RMSE_{CV}$ in the serial gray-box models, *i.e.*, 0.7% and 0.5%, respectively, which means that the LW-PLSR models offset the output error of the serial gray-box models. The combined gray-box models showed the highest prediction accuracy among the white-box models and three types of gray-box models, regardless of the formulation.

### 5.3.2. Applicability to real operating data on commercial scale

Figure 5.3 shows the $RMSE_P$ and $R^2$ boxplots of the white-box model and three types of gray-box models. In formulation A, the median $RMSE_P$ values of the white-box, parallel gray-box, serial gray-box, and combined gray-box models were 1.7%, 1.1%, 1.0%, and 0.8%, respectively. The median $RMSE_P$ values of the combined gray-box models were 53%, 27%, and 20% smaller than those of the white-box, parallel gray-box, and serial gray-box models, respectively. Reflecting these results, the median $R^2$ values of the white-box, parallel gray-box, serial gray-box, and combined gray-box models were 0.68, 0.87, 0.90, and 0.95, respectively. Hence, in the case of formulation A, the combined gray-box model showed the highest prediction accuracy among the three types of gray-box models and met the $RMSE_P$ criterion.

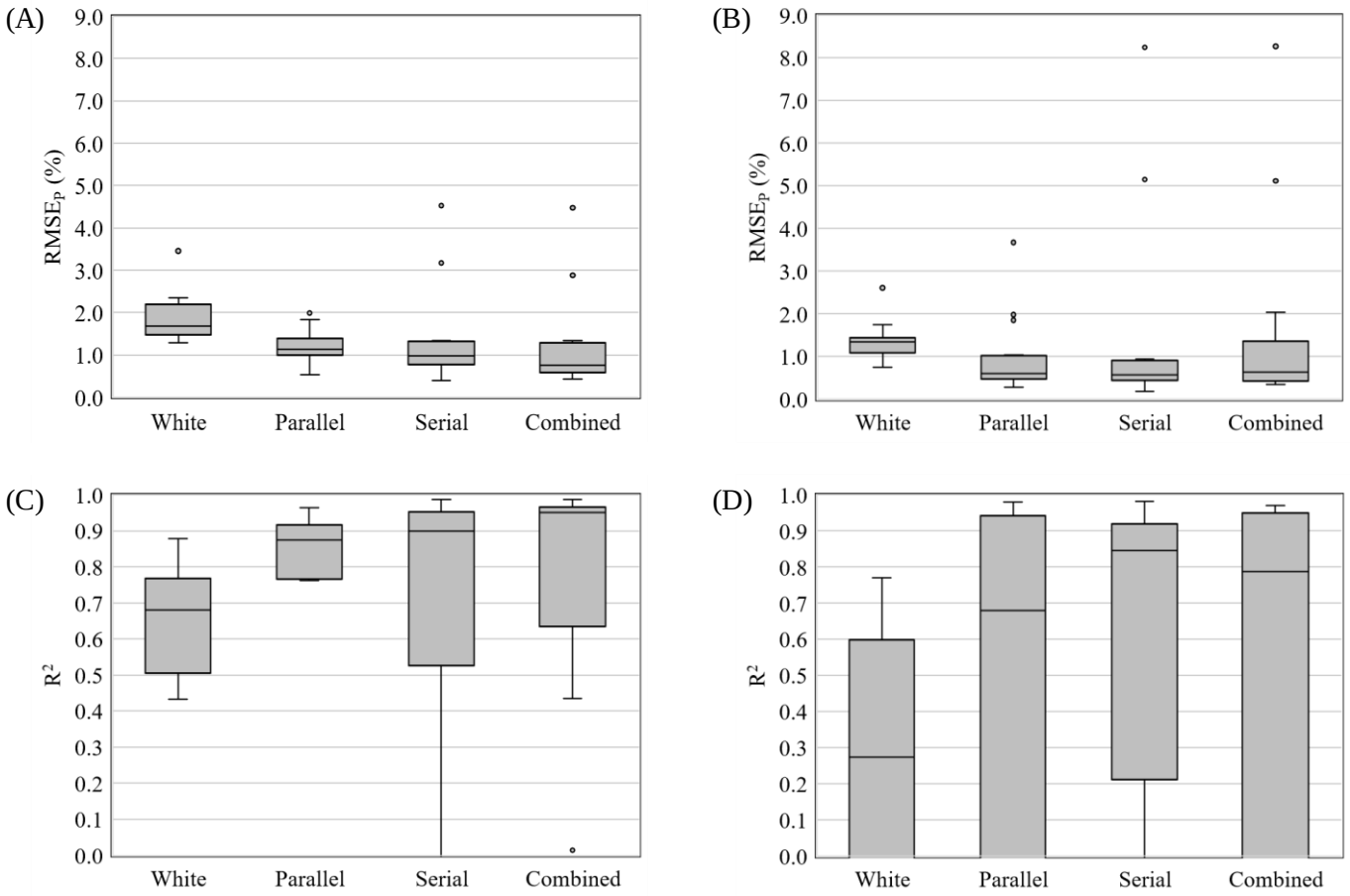


Figure 5.3. Prediction accuracy of white-box, parallel gray-box, serial gray-box, and combined gray-box models in case of (A, C) formulation A and (B, D) formulation B.

(A, B) $RMSE_P$ boxplots, (C, D) $R^2$ boxplots.

On the other hand, in formulation B, the median $RMSE_P$ values of the white-box, parallel gray-box, serial gray-box, and combined gray-box models were 1.3%, 0.6%, 0.6%, and 0.6%, respectively. The median $RMSE_P$ values of the three gray-box models were equivalent and smaller than the $RMSE_P$ criterion. In contrast, the difference between maximum and minimum $RMSE_P$ values except for outliers of the combined gray-box model was more than twice those of the parallel gray-box and serial gray-box models. In other words, the parallel gray-box and serial gray-box models were more robust than the combined gray-box model. Additionally, the median $R^2$ values of the white-box, parallel gray-box, serial gray-box, and combined gray-box models were 0.27, 0.68, 0.85, and 0.79, respectively. Although the $R^2$ variations of all four models were considerable (see Figure 5.3), this is mainly caused by the difference in the reference values range of granule water content, which depends on lot selection. For example, in the case of Trial No. 12 of formulation B (see Table 5.1), the $RMSE_P$ value of the serial gray-box model was 0.7% and smaller than the $RMSE_P$ criterion, whereas the $R^2$ value of the model was -18.0. This is because the range of reference values from 2.5% to 3.0% is too narrow compared to the acceptable output error, *i.e.*, the $RMSE_P$ criterion of 1.0% (refer to Section 5.2.6). Considering an actual process development, these cases where the range of water content is narrow, like Trial No. 12, are unrealistic because the experimental data were acquired with various manufacturing conditions through a process development study using commercial-scale equipment. Therefore, it was concluded that the serial gray-box model was the best estimator of the water content in the case of formulation B.

These results demonstrated that the gray-box models improved the prediction accuracy of the white-box models regardless of the formulation, as shown in Figure 5.4. The best model among the three types of gray-box models differed depending on the

formulations, which means that we should select an accurate and robust model for each formulation. Additionally, this study clarified the point to consider when applying the gray-box models using LW-PLSR to commercial production. As shown in Figure 5.3, in formulation B, the serial gray-box and combined gray-box models showed the extreme outliers of $RMSE_P$, *i.e.*, 8.2% and 8.3%, respectively, which were calculated in Trial No. 8 (see Table 5.1). In both gray-box models, $T^2$ or Q values of more than 90% validation samples exceeded the 99% confidence limit. Reflecting these results, the $RMSE_P$ values of the serial gray-box and combined gray-box models were 8.2% and 8.3%, respectively, and much larger than that of the white-box model, *i.e.*, 1.7%. Thus, we should assess the validity of the LW-PLSR model for a query in terms of two distance criteria, *i.e.*, Hotelling's $T^2$ and Q residual; if the $T^2$ or Q value of the query is higher than the 99% confidence limit, we should consider adopting the white-box model because the LW-PLSR model could deteriorate the prediction accuracy of the white-box model.

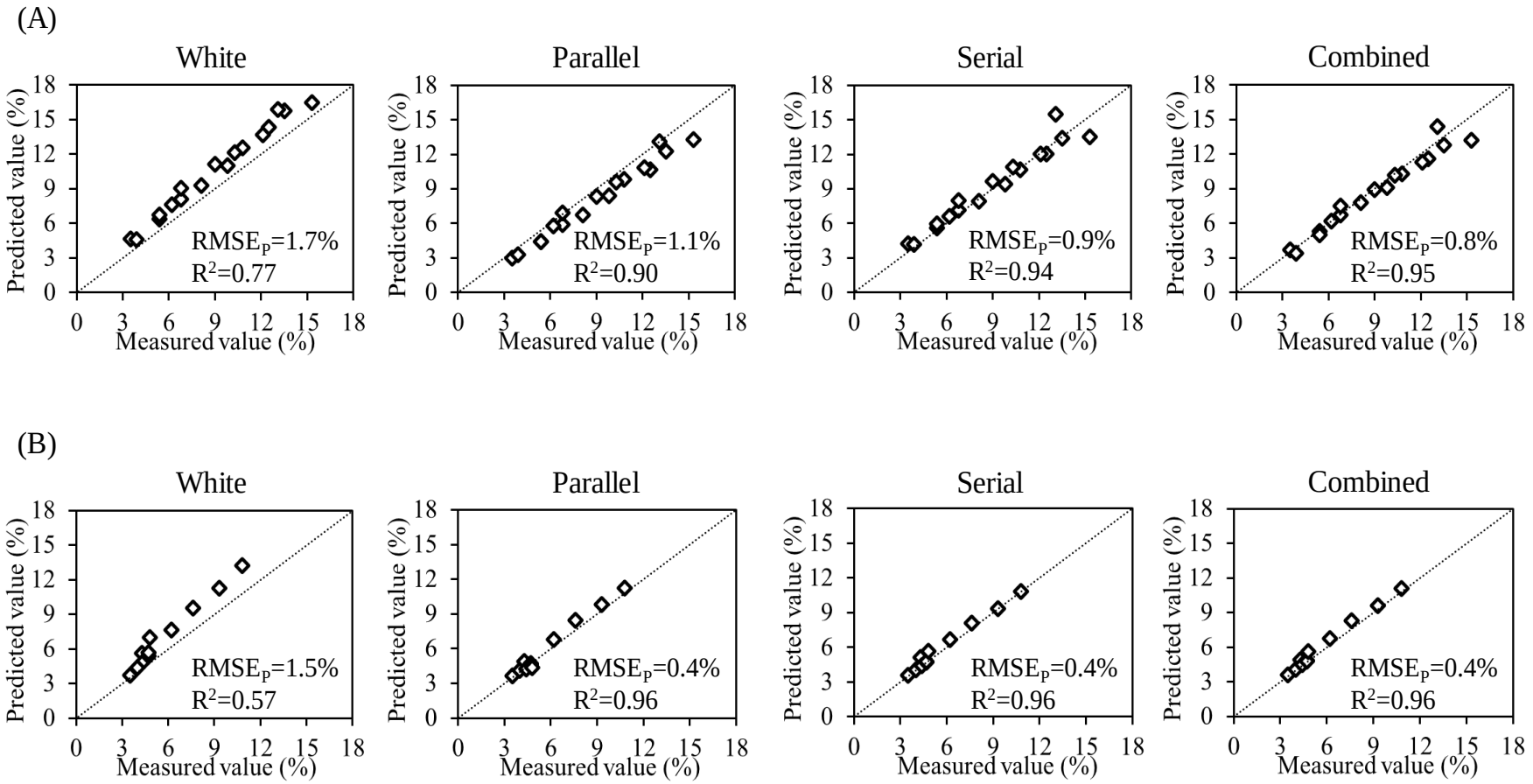


Figure 5.4. Scatter plots of measured vs. predicted water content using white-box, parallel gray-box, serial gray-box, and combined gray-box models, in case of (A) formulation A (Trial No. 6) and (B) formulation B (Trial No. 7).

## 5.4. Summary

In Chapter 5, descriptive and accurate three types of gray-box models for water content monitoring in fluidized bed granulation were developed. These gray-box models, *i.e.*, parallel gray-box, serial gray-box, and combined gray-box models, were constructed by integrating the heat and mass balance model (white-box model) and the LW-PLSR model (black-box model). Their applicability to real operating data on a commercial scale was demonstrated using validation datasets with two formulations. Regardless of the formulation, all three types of gray-box models improved the prediction accuracy of the white-box models. The best model among the three types of gray-box models was different depending on the formulations, which means that we should adopt an accurate and robust model for each formulation. Besides, this study revealed the point to consider when applying the gray-box models using LW-PLSR to commercial production. Whenever a prediction is required for a query, we should assess the validity of the LW-PLSR model for the query based on two distance criteria, *i.e.*, Hotelling's $T^2$ and Q residual. Considering the assessment result, we should judge whether the gray-box soft sensor is suitable for precise process monitoring.

Furthermore, the developed gray-box models require only basic manufacturing information regarding PPs and formulations, which means that they are simple and easy to apply. Therefore, the proposed three types of gray-box models, which balance prediction accuracy and interpretability, are expected to be alternative soft sensors to the white-box and black-box models.

# 6. Conclusion

In the pharmaceutical industry, QbD and PAT have been recently noted to assure pharmaceutical quality at a higher level. Since soft sensors enhance the process understanding, they play a crucial role throughout the pharmaceutical lifecycle. In contrast, the soft sensor's features required at the two key stages, *i.e.*, the technology transfer and the commercial manufacturing, are different.

This thesis developed the following two practical soft sensors to satisfy all the requirements at the technology transfer (scale-free, accurate, and cost-effective) and the commercial manufacturing (descriptive, accurate, and simple) stages, respectively.

- A scale-free, accurate, and cost-effective PP-based black-box soft sensor for technology transfer (Chapters 3 and 4)
- A descriptive, accurate, and simple gray-box soft sensor for commercial manufacturing (Chapter 5)

In Chapter 3, a scale-free, accurate, and cost-effective PP-based black-box model was developed to cope with the manufacturing scale change in the technology transfer stage. The proposed method exploits two key ideas to construct a scale-free PP-based black-box model. First, to accommodate the manufacturing scale change, the critical PPs whose effects on water content are constant among manufacturing scales were selected as input variables. Second, to construct an accurate statistical model, LW-PLSR that can address collinearity and nonlinearity was utilized. The PP-based black-box model was developed using both laboratory-scale and pilot-scale experimental data. The prediction accuracy in the commercial scale was evaluated to align with the situation at the technology transfer

stage. The developed scale-free PP-based black-box model exhibited a high prediction accuracy, which was equivalent to the commonly-used NIRS-based black-box model.

In Chapter 4, further evaluation on the PP-based LW-PLSR model developed in Chapter 3 was conducted. The scale-free PP-based and NIRS-based LW-PLSR models were constructed using different calibration datasets and compared in terms of prediction accuracy and development cost. The development cost was defined as the cost of goods required to build an accurate model of commercial-scale equipment. The number of granulation lots in the calibration datasets was increased by one lot in order of laboratory, pilot, and commercial scale to align with the general stage-up situation from the pharmaceutical development stage to the technology transfer stage. This study revealed that the construction of accurate LW-PLSR models requires the calibration samples with the following two features: 1) located near the validation samples on the subspace spanned by PCs, and 2) having a wide range of variations in PC scores. Furthermore, it was demonstrated that the reduction in cost and mass fraction of API made the proposed PP-based LW-PLSR models more cost-effective than the NIRS-based LW-PLSR models.

In Chapter 5, a descriptive, accurate, and simple gray-box model was developed to enhance continual improvement with an insightful process understanding in the commercial manufacturing stage. The three types of gray-box models, *i.e.*, parallel, serial, and combined gray-box models, were evaluated in terms of prediction accuracy using real operating data on a commercial scale with two formulations. The gray-box models were constructed by integrating the heat and mass balance model (white-box model) and the LW-PLSR model (black-box model). In the serial gray-box models, LW-PLSR models adjusted the fitting parameters of the white-box model depending on the PPs for each query. In the parallel gray-box or combined gray-box models, LW-PLSR models

compensated for the output error of the white-box or serial gray-box models, respectively. The results demonstrated that all three types of gray-box models improved the prediction accuracy of the white-box models regardless of the formulation. Besides, the thesis proposed the assessment method based on Hotelling's $T^2$ and Q residual for gray-box models using LW-PLSR, which contributes decision support to select a gray-box or white-box model.

Throughout the pharmaceutical lifecycle, scale-free and accurate NIRS-based black-box models have been mainly adopted at the expense of a high initial investment cost of the NIR spectrometer and the probe. An alternative method that requires no initial investment has been naturally desired for cost-saving. However, it has been challenging to build an accurate white-box model or PP-based black-box model applicable to commercial scale without using the commercial-scale experimental data. Thus, the successful development of a scale-free, accurate, and cost-effective PP-based black-box soft sensor is a breakthrough in water content monitoring. Furthermore, this thesis proposed a decision support method to select a PP-based or NIRS-based LW-PLSR model considering prediction performance and economic efficiency. The innovative PP-based LW-PLSR model is expected to be an alternative to the existing NIRS-based method, which enhances the implementation of real-time monitoring at the technology transfer stage. Heat and mass balance models are based on the first-principle, and their high interpretability enables industry operators to understand the process in detail. However, non-descriptive NIRS-based black-box models have been extensively utilized for commercial production because heat and mass balance models have a known drawback of prediction accuracy. Hence, the developed descriptive, accurate, and simple three types of gray-box models are valuable to promote industry operators' process understanding for

continual improvement and precise quality control at the commercial manufacturing stage. Moreover, the developed gray-box soft sensors are expected to be useful for feedback or feedforward control because their input variables contain operable PPs critical to the water content, such as spray rate, inlet air volume, and inlet air temperature. In conclusion, the thesis provides novel and practical soft sensors for expanding PAT implementation and enhancing process understanding of the fluidized bed granulation throughout the pharmaceutical lifecycle.

Finally, this thesis discusses a further application of the developed innovative scale-free PP-based LW-PLSR model. This thesis focused on a scaling-up at the technology transfer stage. In contrast, manufacturing site change or addition is also conducted according to the product demand at the commercial manufacturing stage. A NIRS-based monitoring method is available only when the new manufacturing site has the equivalent special instruments, such as the NIR spectrometer and the probe. On the other hand, the developed scale-free PP-based LW-PLSR model could apply to a new fluidized bed granulator with standard instruments, such as a thermometer and hygrometer. Thus, the developed scale-free PP-based LW-PLSR model is practical and easy to apply, and it will be a powerful tool to assure consistency in product quality before and after the change of manufacturing scale or site. Besides, a scale-free PP-based LW-PLSR model could potentially estimate other granule properties, such as particle size, depending on the formulations. The relationship between granule particle size and water content must be constant among manufacturing scales to realize scale-free prediction of granule particle size. When the mass fraction of binder in the formulation is insufficient, scaling-up usually makes the particle size of the fragile granule smaller, which is a difficult case to apply scale-free prediction using the PP-based LW-PLSR model. Furthermore, the

concept of the scale-free PP-based LW-PLSR model could be applied to water content monitoring in other manufacturing processes, such as film coating. Therefore, further evaluation for applying the proposed concept of scale-free PP-based black-box soft sensors to various MAs and manufacturing processes could lead to more cost-effective PAT implementation, which contributes to the higher-level assurance of pharmaceutical quality.

## Acknowledgements

This dissertation has been completed with a great deal of support and assistance.

First, I would like to express my deepest appreciation to Prof. Dr. Manabu Kano for this precious opportunity to accomplish my doctoral thesis under his supervision. His insightful guidance always made my thinking sharpen and brought my research to a higher level.

I am also heartily grateful to Prof. Dr. Shin Ishii and Prof. Dr. Hidetoshi Shimodaira for their invaluable discussion. The constructive advice and comments given by them has been a great help in writing this dissertation.

I would like to acknowledge Prof. Dr. Hirokazu Sugiyama from The University of Tokyo for his intelligent feedback on my research. Discussions with him have been meaningful and made my research achievement great.

Special thanks go to Dr. Shuichi Tanabe, Dr. Takuya Miyano, and Dr. Hiroshi Nakagawa from Daiichi Sankyo Co., Ltd. for their valuable cooperation in my research and considerable encouragement. I am also very thankful to my seniors and colleagues in Daiichi Sankyo Co., Ltd. for making my doctoral thesis possible with the generous support and consideration.

Finally, I would like to show my heartfelt appreciation to my family, Keiko and Yuta, for their sincere support, understanding, and encouragement throughout my study.

# Abbreviations

| | | |
|---|---|---|
| API | : | Active Pharmaceutical Ingredient |
| CMA | : | Critical Material Attribute |
| CPP | : | Critical Process Parameter |
| CQA | : | Critical Quality Attribute |
| EMA | : | European Medicines Agency |
| FDA | : | Food and Drug Administration |
| ICH | : | International Conference on Harmonisation of Technical Requirements for Registration of Pharmaceuticals for Human Use |
| LOD | : | Loss on Drying |
| LOOCV | : | Leave-One-Out Cross Validation |
| LW-PLSR | : | Locally Weighted Partial Least Squares Regression |
| MA | : | Material Attribute |
| NCSC | : | Nearest Correlation Spectral Clustering |
| NIR | : | Near-Infrared |
| NIRS | : | Near-Infrared Spectroscopy |
| PAT | : | Process Analytical Technology |
| PBM | : | Population Balance Model |
| PC | : | Principal Component |
| PCA | : | Principal Component Analysis |
| PLSR | : | Partial Least Squares Regression |
| PP | : | Process Parameter |
| QbD | : | Quality by Design |

| | | |
|---|---|---|
| QTPP | : | Quality Target Product Profile |
| R | : | Correlation Coefficient |
| $R^2$ | : | Coefficient of Determination |
| $RMSE_C$ | : | Root Mean Square Error of Calibration |
| $RMSE_{CV}$ | : | Root Mean Square Error of Cross Validation |
| $RMSE_P$ | : | Root Mean Square Error of Prediction |
| RTRT | : | Real-Time Release Testing |
| SFD | : | Spectral Fluctuation Dividing |
| SNV | : | Standard Normal Variate |
| VIP | : | Variable Importance in the Projection |

# References


1) Kristensen H.G., Schaefer T. Granulation: A Review on Pharmaceutical Wet-Granulation. *Drug. Dev. Ind. Pharm.* (1987) 13, 803-872.

2) Suzuki Y., Kato T., Kikkawa Y., Suzuki T., Wakiyama N., Terada K. Scale-up and blender change model for the pharmaceutical lubricated mixing process. *Powder Technol.* (2015) 280, 113-118.

3) Knop K., Kleinebudde P. PAT-tools for process control in pharmaceutical film coating applications. *Int. J. Pharm.* (2013) 457, 527-536.

4) Rambali B., Baert L., Massart D.L. Using experimental design to optimize the process parameters in fluidized bed granulation on a semi-full scale. *Int. J. Pharm.* (2001) 220, 149-160.

5) Mehta A.M. Scale-up considerations in the fluid-bed process for controlled-release products. *Pharm. Tech.* (1988) 12, 46-52.

6) Hartung A., Knoell M., Schmidt U., Langguth P. Role of continuous moisture profile monitoring by inline NIR spectroscopy during fluid bed granulation of an Enalapril formulation. *Drug. Dev. Ind. Pharm.* (2011) 37, 274-280.

7) Burggraeve A., Monteyne T., Vervaet C., Remon J.P., Beer T.D. Process analytical tools for monitoring, understanding, and control of pharmaceutical fluidized bed granulation: A review. *Eur. J. Pharm. Biopharm.* (2013) 83, 2-15.

8) Rambali B., Baert L., Massart D.L. Scaling up of the fluidized bed granulation process. *Int. J. Pharm.* (2003) 252, 197-206.

9) Watano S., Fukushima T., Miyanami K. Heat Transfer and Granule Growth Rate in Fluidized Bed Granulation. *Chem. Pharm. Bull.* (1996) 44, 572-576.

10) Tanino T., Aoki Y., Yoshida T., Sumi Y., Mizuta T. Granule Size Control and Scale-up Strategy in Fluidized-bed Granulation. *PDA J. GMP valid. Japan.* (2000) 2, 19-27.

11) Heinrich S., Henneberg M., Peglow M., Drechsler J., Mörl L. Fluidized Bed Spray Granulation: Analysis of Heat and Mass Transfers and Dynamic Particle Populations. *Braz. J. Chem. Eng.* (2005) 22, 181-194.

12) Hu X., Cunningham J., Winstead D. Understanding and Predicting Bed Humidity in Fluidized Bed Granulation. *J. Pharm. Sci.* (2008) 97, 1564-1577.

13) Wang H.G., Senior P.R., Mann R., Yang W.Q. Online measurement and control of solids moisture in fluidised bed dryers. *Chem. Eng. Sci.* (2009) 64, 2893-2902.

14) Ochsenbein D.R., Billups M., Hong B., Schäfer E., Marchut A.J., Lyngberg O.K., Industrial application of heat- and mass balance model for fluid-bed granulation for technology transfer and design space exploration. *Int. J. Pharm. X*, (2019) 1, 100028.

15) Randolph A.D., Larson M.A. Theory of Particulate Processes. *Academic Press, New York* (1988).

16) Hounslow M.J., Ryall R.L., Marshall V.R. A discretized population balance model for nucleation, growth and aggregation. *AIChE J.* (1988) 34, 1821-1832.

17) Ramkrishna D. Population Balances: Theory and Applications to Particulate Systems in Engineering. *Academic Press, San Diego* (2000).

18) Vreman A.W., Van Lare C.E., Hounslow M.J. A basic population balance model for fluid bed spray granulation. *Chem. Eng. Sci.* (2009) 64, 4389-4398.

19) FDA, 2004. Guidance for industry, PAT – A framework for innovative pharmaceutical development, manufacturing, and quality assurance.

20) EMA, 2006. Quality risk management (International Conference on Harmonisation

of Technical Requirements for Registration of Pharmaceuticals for Human Use Q9).

21) EMA, 2008. Note for guidance on pharmaceutical quality system (International Conference on Harmonisation of Technical Requirements for Registration of Pharmaceuticals for Human Use Q10).

22) EMA, 2009. Note for guidance on pharmaceutical development.

23) ICH, 2005. ICH Harmonised Tripartite Guideline-Quality Risk Management Q9.

24) ICH, 2005. ICH Harmonised Tripartite Guideline-Pharmaceutical Quality System Q10.

25) ICH, 2009. ICH Harmonised Tripartite Guideline-Pharmaceutical Development Q8 (R2).

26) Yu L.X. Pharmaceutical Quality by Design: Product and Process Development, Understanding, and Control. *Pharm. Res.* (2008) 25, 781-791.

27) Huang J., Kaul G., Cai C., Chatlapalli R., Hernandez-Abad P., Ghosh K., Nagi A. Quality by design case study: An integrated multivariate approach to drug product and process development. *Int. J. Pharm.* (2009) 382, 23-32.

28) Acharya S., Kam M. Evidence Combination for Hard and Soft Sensor Data Fusion. *14th International Conference on Information Fusion* (2011) 1008-1015.

29) Kadlec P., Gabrys B., Strandt S., Data-driven Soft Sensors in the process industry. *Comput. Chem. Eng.* (2009) 33, 795-814.

30) Kano M., Fujiwara K. Virtual Sensing Technology in Process Industries: Trends and Challenges Revealed by Recent Industrial Applications. *J. Chem. Eng. Japan*. (2013) 46, 1-17.

31) Frake P., Greenhalgh D., Grierson S.M., Hempenstall J.M., Rudd D.R. Process control and end-point determination of a fluid bed granulation by application of near

infra-red spectroscopy. *Int. J. Pharm.* (1997) 151, 75-80.

32) Rantanen J., Antikainen O., Mannermaa J.P., Yliruusi J. Use of the Near-Infrared Reflectance Method for Measurement of Moisture Content During Granulation. *Pharm. Dev. Technol.* (2000) 5, 209-217.

33) Green R.L., Thurau G., Pixley N.C., Mateos A., Reed R.A., Higgins J.P. In-Line Monitoring of Moisture Content in Fluid Bed Dryers Using Near-IR Spectroscopy with Consideration of Sampling Effects on Method Accuracy. *Anal. Chem.* (2005) 77, 4515-4522.

34) Alcalà M., Blanco M., Bautista M., González J.M. On-Line Monitoring of A Granulation Process By NIR Spectroscopy. *J. Pharm. Sci.* (2010) 99, 336-345.

35) Berntsson O., Danielsson L.G., Lagerholm B., Folestad S. Quantitative in-line monitoring of powder blending by near infrared reflection spectroscopy. *Powder Technol.* (2002) 123, 185-193.

36) Wu H., Khan M.A. Quality-by-design (QbD): an integrated approach for evaluation of powder blending process kinetics and determination of powder blending endpoint. *J. Pharm. Sci.* (2009) 98, 2784-2798.

37) Porfire A., Rus L., Vonica A.L., Tomuta I. High-throughput NIR-chemometric methods for determination of drug content and pharmaceutical properties of indapamide powder blends for tabletting. *J. Pharm. Biomed. Anal.* (2012) 70, 301-309.

38) Chalus P., Roggo Y., Walter S., Ulmschneider M. Near-infrared determination of active substance content in intact low-dosage tablets. *Talanta* (2005) 66, 1294-1302.

39) Kirsch J.D., Drennen J.K. Determination of film-coated tablet parameters by near-infrared spectroscopy. *J. Pharm. Biomed. Anal.* (1995) 13, 1273-1281.

40) Kirsch J.D., Drennen J.K. Near-infrared spectroscopic monitoring of the film coating process. *Pharm. Res.* (1996) 13, 234-237.

41) Reich G. Near-infrared spectroscopy and imaging: Basic principles and pharmaceutical applications. *Adv. Drug Deliv. Rev.* (2005) 57, 1109-1143.

42) Roggo Y., Chalus P., Maurer L., Lema-Martinez C., Edmond A., Jent N. A review of near infrared spectroscopy and chemometrics in pharmaceutical technologies. *J. Pharm. Biomed. Anal*. (2007) 44, 683-700.

43) Jamrógiewicz M. Application of the near-infrared spectroscopy in the pharmaceutical technology. *J. Pharm. Biomed. Anal*. (2012) 66, 1-10.

44) Ahmad I., Kano M., Hasebe S., Kitada H., Murata N. Gray-box modeling for prediction and control of molten steel temperature in tundish. *J. Process Control* (2014) 24, 375-382.

45) Ahmad I., Ayub A., Kano M., Cheema I.I. Gray-box Soft Sensors in Process Industry: Current Practice, and Future Prospects in Era of Big Data. *Processes* (2020) 8, 243.

46) Lipsanen T., Närvänen T., Räikkönen H., Antikainen O., Yliruusi J. Particle Size, Moisture, and Fluidization Variations Described by Indirect In-line Physical Measurements of Fluid Bed Granulation. *AAPS PharmSciTech* (2008) 9, 1070-1077.

47) Närvänen T., Antikainen O., Yliruusi J. Predicting Particle Size During Fluid Bed Granulation Using Process Measurement Data. *AAPS PharmSciTech* (2009) 10, 1268-1275.

48) Wold S., Sjöström M., Eriksson L. PLS-regression: a basic tool of chemometrics. *Chemometr. Intell. Lab. Syst.* (2001) 58, 109-130.

49) Yoshizaki R., Kano M., Tanabe S., Miyano T. Process Parameter Optimization based

on LW-PLS in Pharmaceutical Granulation Process. *IFAC-PapersOnLine* (2015) 48, 303-308.

50) Liu H., O'Connor T., Lee S., Yoon S. A process optimization strategy of a pulsed-spray fluidized bed granulation process based on predictive three-stage population balance model. *Powder Technol.* (2018) 327, 188-200.

51) Kim S., Kano M., Nakagawa H., Hasebe S. Estimation of active pharmaceutical ingredients content using locally weighted partial least squares and statistical wavelength selection. *Int. J. Pharm.* (2011) 421, 269-274.

52) Johansen T.A., Foss B.A. Representing and Learning Unmodeled Dynamics with Neural Network Memories. *American Control Conference* (1992) 3037-3043.

53) Psichogios D.C., Ungar L.H. A Hybrid Neural Network-First Principles Approach to Process Modeling. *AIChE J.* (1992) 38, 1499-1511.

54) Jackson J.E., Mudholkar G.S. Control Procedures for Residuals Associated With Principal Component Analysis. *Technometrics* (1979) 21, 341-349.

55) Barnes R.J., Dhanoa M.S., Lister S.J. Standard Normal Variate Transformation and De-trending of Near-Infrared Diffuse Reflectance Spectra. *Appl. Spectrosc.* (1989) 43, 772-777.

56) Miyano T., Fujiwara K., Kano M., Tanabe H., Nakagawa H., Watanabe T., Minami H. Efficient wavenumber selection based on spectral fluctuation dividing and correlation-based clustering for calibration modeling. *Chemom. Intell. Lab. Syst.* (2015) 148, 85-94.

57) Nakagawa H., Tajima T., Kano M., Kim S., Hasebe S., Suzuki T., Nakagami H. Evaluation of Infrared-Reflection Absorption Spectroscopy Measurement and Locally Weighted Partial Least-Squares for Rapid Analysis of Residual Drug

Substances in Cleaning Processes. *Anal. Chem.* (2012) 84, 3820-3826.

58) Nakagawa H., Kano M., Hasebe S., Miyano T., Watanabe T., Wakiyama N. Verification of model development technique for NIR-based real-time monitoring of ingredient concentration during blending. *Int. J. Pharm.* (2014) 471, 264-275.

59) Kim S., Okajima R., Kano M., Hasebe S. Development of soft-sensor using locally weighted PLS with adaptive similarity measure. *Chemometr. Intell. Lab. Syst.* (2013) 124, 43-49.

60) Hirai T., Kano M. Adaptive Virtual Metrology Design for Semiconductor Dry Etching Process Through Locally Weighted Partial Least Squares. *IEEE Trans. Semicond. Manuf.* (2015) 28, 137-144.

61) Kamohara H., Takinami A., Takeda M., Kano M., Hasebe S., Hashimoto I. Product Quality Estimation and Operating Condition Monitoring for Industrial Ethylene Fractionator. *J. Chem. Eng. Japan* (2004) 37, 422-428.

62) Muteki K., Yamamoto K., Reid G.L., Krishnan M. De-risking Scale-up of a High Shear Wet Granulation Process Using Latent Variable Modeling and Near-Infrared Spectroscopy. *J. Pharm. Innov.* (2011) 6, 142-156.

63) Chong I.G., Jun C.H. Performance of some variable selection methods when multicollinearity is present. *Chemometr. Intell. Lab. Syst.* (2005) 78, 103-112.

64) Cogdill R.P., Knight T.P., Anderson C.A., Drennen III J.K. The Financial Re turns on Investments in Process Analytical Technology and Lean Manufactur ing: Benchmarks and Case Study. *J. Pharm. Innov.* (2007) 2, 38-50.

65) Wold S., Esbensen K., Geladi P. Principal Component Analysis. *Chemom. In tell. Lab. Syst.* (1987) 2, 37-52.

66) Tanabe S., Nakagawa H., Watanabe T., Minami H., Kano M., Urbanetz N.A. Setting

the process parameters for the coating process in order to assure tablet appearance based on multivariate analysis of prior data. *Int. J. Pharm.* (2016) 511, 341-350.