**Using Machine Learning Tools for Residual Moisture monitoring in Freeze-Drying processes for Pharmaceutical Applications**

Ambra Massei1,2, Serena Bobba1,2, Nunzio Zinfollino2, Davide Fissore1\*

*1 Dipartimento di Scienza Applicata e Tecnologia, Politecnico di Torino, corso Duca degli Abruzzi 24, 10129 Torino*

*2 Biotech Pharmaceutical Development Department, Merck Serono SpA, via Luigi Einaudi 11, 00012 Guidonia Montecelio (Roma)*

*\*Corresponding author E-Mail:* [*davide.fissore@polito.it*](mailto:davide.fissore@polito.it) *(D. Fissore)*

**1.Introduction**

Freeze-drying is a crucial step in many drug manufacturing processes as it provides long-term stability to formulations containing an active pharmaceutical ingredient (API). Approximately 46% of biopharmaceutical dosage forms approved by Food and Drug Administration (FDA) in the period 2003-2011 are prepared by freeze-drying.[1]

A typical freeze-drying process starts with a freezing phase, in which most of the water, the *free water*, crystallizes as ice. A fraction of water, the *bound water*, is incorporated into the frozen product and remains unfrozen. After freezing, there is the primary drying phase, in which pressure is lowered in the chamber to promote sublimation of the ice. This leads to the formation of a porous solid structure, the cake, through which water vapor escapes from the product. Finally, secondary drying takes place, in which the temperature is increased to promote the desorption of bound water and achieve the target value of residual moisture in the final product.[1]

Pharmaceutical companies must meet standards imposed by regulatory agencies, so final products must meet Critical Quality Attributes. Among all, the residual moisture (RM) plays an important role because it is closely related to the stability of the active ingredient. At the end of a freeze-drying cycle, the suggested moisture range is approximately 3-6%. So, it is necessary to monitor the residual amount of water in the final product. The residual moisture is generally investigated by Karl Fisher (KF) titration. This method is time-consuming, destructive and can therefore only be applied to a part of a production batch. Near-Infrared Spectroscopy (NIR Spectroscopy), on the other hand, is a fast, non-destructive method that can be applied to many samples. In NIR spectroscopy the sample is irradiated by a beam of NIR light and some of this NIR light is absorbed by the molecules, bringing them to a higher vibrational state. So, NIR spectroscopy is about studying the vibrational transitions in molecules, due to the absorbance of the system in the NIR region, i.e. in a region of the electromagnetic spectrum corresponding to 14300 - 4000 cm-1.[2]

This is in line with the Process Analytical Technology (PAT), a concept proposed by the U.S. FDA in 2002 that is expected to lie at the basis of the pharmaceutical “Good Manufacturing Practice” rules.[3] This approach introduces the concept of Quality-by-Design (QbD) according to which the quality should be built into products. This is completely different from the Quality-by-Testing approach in which the quality is tested into products.[4]

In this framework, NIR spectroscopy was used as method to measure RM in freeze-dried products. Here, a *robust* model was developed using *machine learning* tools: the goal of the model is to estimate RM from NIR spectra. The aim was to find a model able to predict the RM values of products also different from those used in the training phase of the model. In particular, the cases of a different percentage of solid fraction, a different excipient and the presence of another component in the product were considered. Great attention was paid in choosing the wavelength range over which to build the model. Looking at the spectra of the various sample data sets, it was chosen to focus on the range that contains 5150 cm-1, the water peak. In fact, water is a component that all formulations have in common. As a result, it was thought that, to obtain a robust model, it might be effective to focus on a narrow range of wavelengths that encompassed the water peak. In this way, specific peaks in product characteristics have a lower impact on the spectra. In addition, it would not become necessary to include all products in the calibration step, allowing for a reduction of the experimental effort for model development.

**2. Methods**

Some freeze-drying cycles were carried out to get the samples used for model development. In the first part of the study, a sucrose 6%w solution, freeze-dried into 2R glass vial (Nuova Ompi, Piombino Dese, Italy) with a filling volume of 1 mL, was considered for samples preparation. Vials were placed according to a honeycomb layout surrounded by metal frames, in direct contact with the freeze-dried shelves, and processed in a lab-scale freeze-drier (Lyostar3, SP Scientific, Warminster, USA) in the laboratories of the Guidonia Montecelio (Italy) site of Merck Serono SpA.[5] In order to get a RM in the range 1-5% some vials were humidified on purpose after the freeze-drying step. Sample sets having specific different features were necessary to test the robustness of the model. For this reason, the following samples were produced:

* 25 samples of sucrose 9%w  (sample set S9), thus having a solid fraction higher than S6;
* 22 samples of sucrose 3%w (sample set S3), thus having a solid fraction lower than S6;
* 35 samples of trehalose 6%w  (sample set T6), thus a different amorphous excipient;
* 20 samples of sucrose 6%w – arginine 1%w (sample set SA14), thus 14.3% of arginine;
* 10 samples of sucrose 6%w – arginine 0.5%w (sample set SA7), thus 7.7% of arginine.

All RM values less than 1% were removed because they exceeded the lower limit of the experimental technique used (KF). So at the end the number of samples for each case was lower, as it can be seen in Table 1.

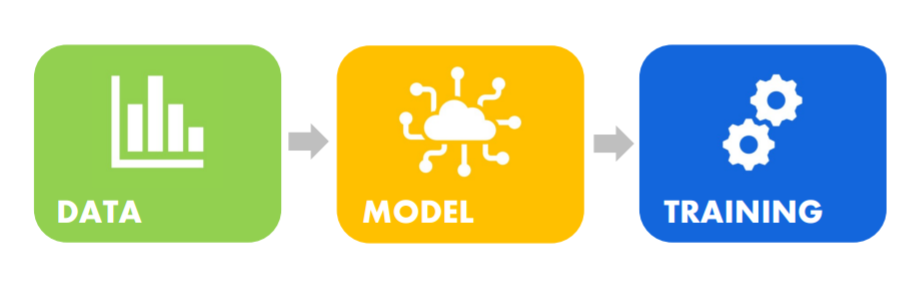
**Table 1:** Number of samples for each data set

|  |  |
| --- | --- |
| *Data set* | *N° Samples* |
| *S6-V* | 96 |
| *S6-H* | 96 |
| *S3* | 22 |
| *S9* | 8 |
| *SA7* | 8 |
| *SA14* | 5 |
| *T6* | 26 |

The samples listed above were all analyzed through a Fourier Transform NIR spectrometer (Antaris MX FT-NIR, Thermo Fischer Scientific, Waltham, USA), equipped with an InGaAs detector and a halogen NIR source. Pre-processing of the spectra was carried out by using a standard normal variate (SNV) approach, consisting in subtracting the mean value of the spectrum from all its points, and dividing for its standard deviation. It is possible to have get more in-depth details on the conduct of the trials in the paper by Bobba et al.[5]

*Machine Learning Tools*

Machine Learning techniques differ from traditional algorithms because they also have the ability to learn as well as apply pre-programmed decisions. Traditional software receives input data and user-written code and generates an output. Machine Learning algorithms, on the other hand, are able to find the functional relationship that binds the input data with the desired output. Models are obtained from a set of data, as shown in Fig. 1. Their formulation does not require a priori knowledge of the physics governing the system or of the relationships linking the input and output variables, although this may be useful.



**Figure 1**: Building a model using machine learning tools

The data set contains all the information necessary to obtain the model: in this case, they are the matrix of the spectra *X* and the vector of residual moisture *y*. The rows of the *X*-matrix are the observations and the columns represent the absorbance corresponding to a specific wavelength.

Building a machine learning model requires a data splitting into two different set:

1. Training set: the model processes the spectra (input) with the RM values obtained with KF (output);
2. Test set: the trained model is used to determine RM in samples not used in the calibration set. The values obtained, RM predicted by the model, are compared with those measured with KF.

For processing NIR spectra and modeling, scripts were written using Spyder (Python 3.9), exploiting the module *Scikit-Learn*. In particular, a multivariate linear regression problem was solved by using the “*LinearRegression*” tool. For regression, the general prediction formula for a multivariate linear model looks as:

|  |  |  |
| --- | --- | --- |
|  |  |  |

Here, to denotes the input variables (*p* is the number of features); *i* refers to the rows of the matrix *X*; and are parameters of the model that are learned, and is the prediction the model makes.[6]

In order to evaluate the performance of the model, the Root Mean Square Error (RMSE) and the Coefficient of determination, , were adopted. They give an idea of how much error the system typically makes in its predictions. So, linear regression finds the parameters and that minimize the RMSE between predictions and the true regression targets, *y*, on the training set. Equations 2 and 3 show the mathematical formulas to compute the RMSE and .

|  |  |  |
| --- | --- | --- |
|  |  |  |
|  |  |  |

The total number of observations and the mean value are designed respectively with and . The lower the RMSE value, the better the model performance.[7] The closer the value of R2 is to 1, the more the points are aligned to the bisector in the parity plot. Parity plots were also made, which correlate the RM values calculated by the model (*y*-axis) with the RM values measured by the KF (*x*-axis). It has been useful to report also the diagrams that correlate the absolute error with the measured RM values considering an *intrinsic error* of the experimental technique of 0.3%.[3]

*Model Development*

Data set S6-V was used for developing two models:

1. Model SR, based on a regression on a small wavelength range, ;
2. Model WR, based on a regression on a large wavelength range, .

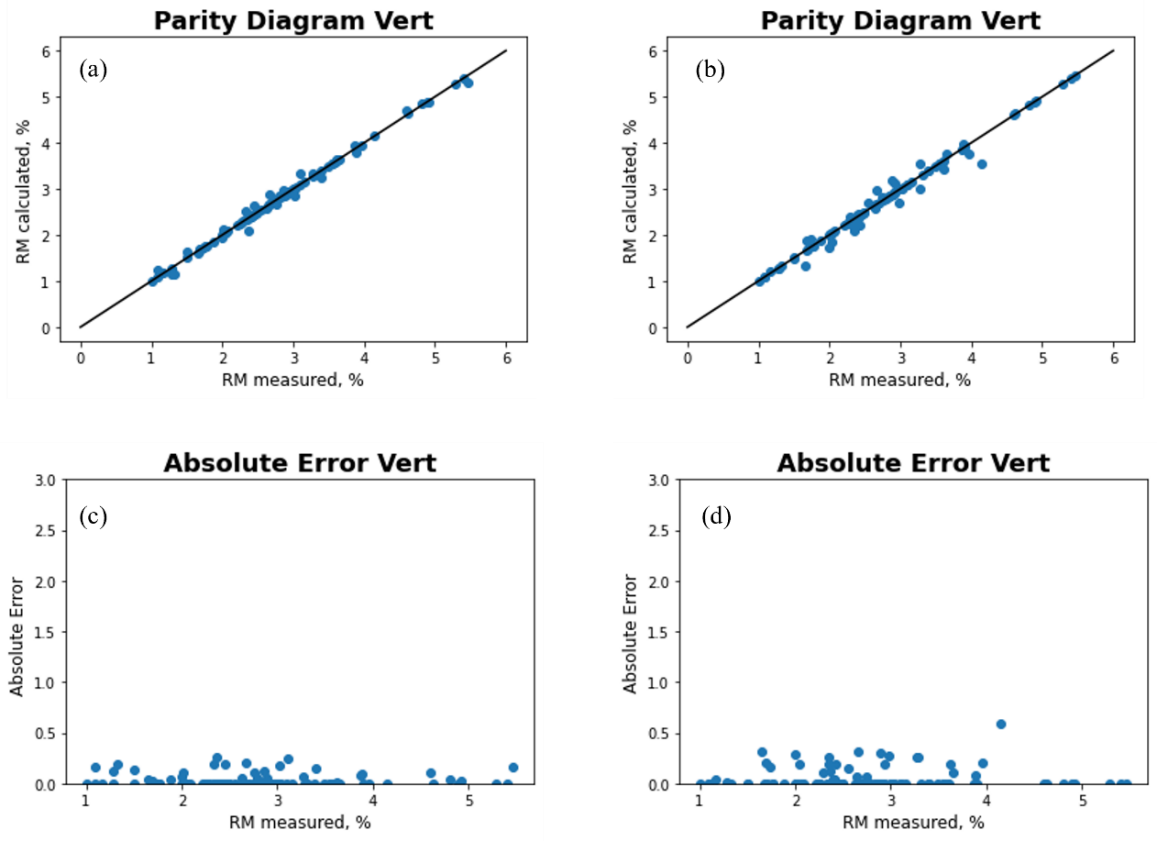
The data set was split in two sets, in order to perform the training with a set and the validation with the other. The size of the training set was varied, from 30% to 70%, to find the minimum that would lead to an acceptable result in order to reduce the experimental effort. Then, the other data sets were used to perform the external validation of the model and challenge its robustness. Also a global model (GM) was developed by considering all the products previously listed as a single data set used also for the training phase. Since model GM was calibrated with different products, it was assumed to provide the best performances in prediction.

**3. Results and discussion**

Model SR, WR and GM are compared in the followings to establish which is the most suitable method for predicting the RM content. The most significant results obtained are reported. Both model WR and SR appeared to have comparable performance and seemed to be accurate enough to predict the RM content of formulations not included in the calibration set. In fact, the RMSE and values in the two cases are almost similar, as it can be seen in Table 2.

**Table 2**: Values of RMSE and R2 for each data set

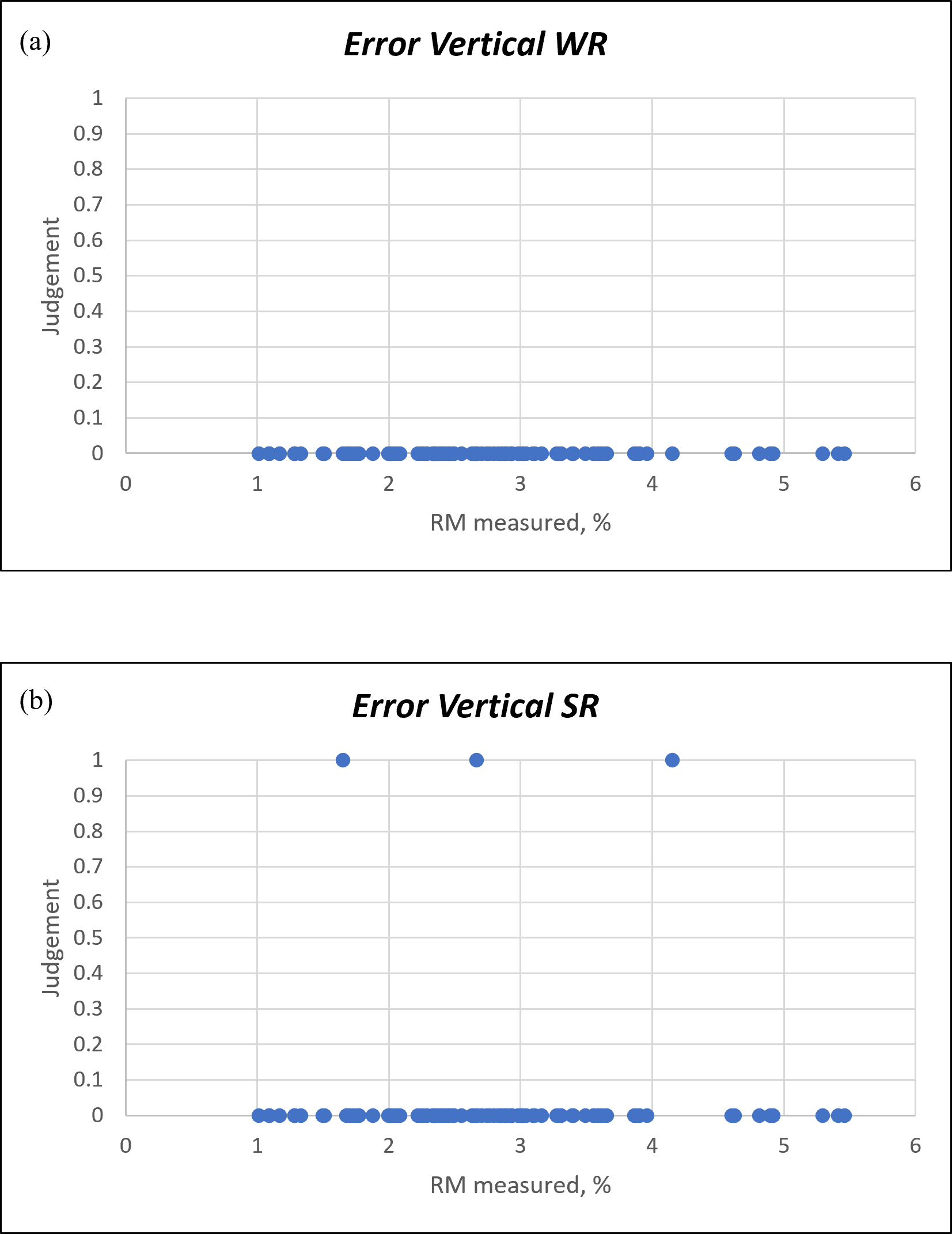
|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| *Data set* | *R2 SR* | *RMSE SR* | *R2 WR* | *RMSE WR* |
| *S6-V* | 0.993 | 0.119 | 0.995 | 0.074 |
| *S6-H* | 0.971 | 0.47 | 0.966 | 0.294 |
| *S9* | 0.997 | 0.281 | 0.992 | 0.248 |



**Figure 2**: Correlation plots between RM (%) measured vs RM (%) calculated and Absolute error plots, obtained from the data set S6-V and processed by: (a),(c) model WR; (b),(d) model SR.

The correlation plots confirm the conclusions obtained from the RMSE and R2 values shown in Table 2. In Fig. 2a and 2b it is possible to notice that the observations are distributed similarly along the bisector in the two cases, index of a good performance of the models.

To evaluate the percentage of success and failure of the obtained model, a vector called "*judgement*" was constructed. The intrinsic error in the KF experimental technique is 0.3%. The absolute error was calculated as the difference in absolute value between the value predicted by the model and the value measured by the KF . In the vector the value 1 occurs when the difference is greater than 0.3%; the value 0 if the difference is less than 0.3%. Figure 3 shows the trends obtained for the S6-V data set with both models. The number of times that in the diagram there is the value 1 corresponds to the number of times that the model exceeds the experimental error and, therefore, fails. The WR model had a good prediction of 100% as all observations were within the range given by the experimental error. In fact, all values in Fig. 3a are equal to 0. In contrast, the SR model was able to correctly predict 96.88% of the observations. In fact, 3 values equal to 1 are observed in Fig. 3b.

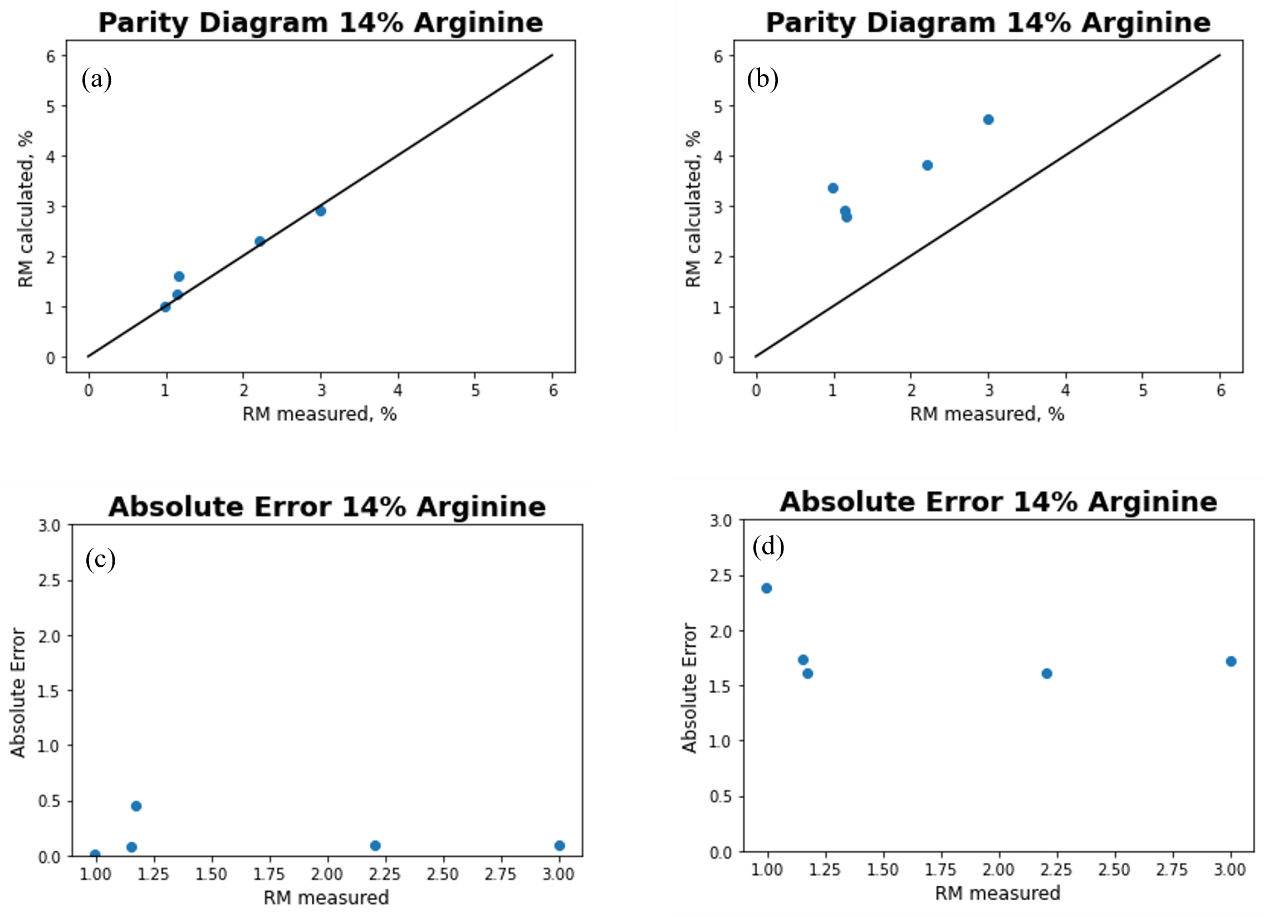


**Figure 3**: Success rate of the algorithm obtained with the S6-V data set: (a) model WR; (b) model SR.

The external validation of model WR and SR with S6-H spectra also yielded in very good performance parameters. Therefore, model SR and WR could be equally applied to spectra collected with vertical and horizontal layout. The fact that the spectra collected in vertical layout and in horizontal layout were similar was a confirmation of what Bobba et al. found by applying the PLS regression. This means that the spectra collected from two different spots of the cake are basically equivalent.[5]

The effect of a different solid fraction was evaluated by means of data set S9. The external validation of model WR with data set S9 was less accurate at lower RM, where the distribution of the observations in the correlation plot is wider. For larger RM values, however, the points lie almost perfectly aligned to the bisector with both models. In order to confirm the robustness of the obtained model, however, it would be appropriate to expand the data set with new experimental tests. In fact, only 8 data were made available for the S9 data set.

The effect of adding an amino acid was studied with 7% and 14% sucrose-arginine mixtures. The case of 14% arginine concentration is reported in Fig. 4.



**Figure 4**: Correlation plots for sucrose-arginine mixture and Absolute Error plots processed by: (a),(c) model WR; (b),(d) model SR.

Model SR turned out to have the poorest performances in comparison with model WR to predict the residual moisture in a sample where an amino acid is added to the excipient considered in the training phase. In fact, in Fig. 4b it can be observed that the observations are much more spread from the bisector than those processed with the SR model (Fig. 4a). These findings are confirmed also by the values of R2 and RMSE for both model, shown in Table 3.

**Table 3**: Values of RMSE and R2 for sucrose-arginine mixture

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| *Data set* | *R2 SR* | *RMSE SR* | *R2 WR* | *RMSE WR* |
| *SA14* | 0.863 | 1.837 | 0.949 | 0.211 |

These differences could be explained by looking at the spectra of the sucrose-arginine mixture. Significant differences can be seen in the region analyzed by the SR model compared to the 6% sucrose mixture used to calibrate the model.

Both the models were also tested using a different excipient, a 6% trehalose solution. The failure of the external validation applied to this data set appeared clear. The correlation plots showed a muddled distribution. The poor performance parameters confirm these findings. Due to the limited number of samples available, a more extended calibration might be advisable. Future work may try to use a nonlinear model, such as a neural network, which may have better performance in predicting an excipient different from the one used in the calibration step.

**4. Conclusions**

Machine learning tools were applied to develop a model suitable for the prediction of the RM content of freeze-dried products, focusing on its robustness and suitability for different products not involved in the training procedure.

Various tests have been done to find the minimum size of the training set that generates acceptable results, thus reducing the experimental effort required. Both models, SR and WR, were found to be very accurate in describing samples with high percentage of solid (S9). In contrast, they showed poor performance in predicting samples made from a different excipient (T6). Model WR was more accurate in predicting samples made from sucrose-arginine mixtures than the SR model. It has to be remarked that the accuracy of this method is strongly dependent on the accuracy of the KF titration used for model training.

A global model GM, calibrated with all products, gave basically the same results as models WR and SR, except for the samples containing trehalose, which were better described.

Future work may focus on applying more complex, nonlinear algorithms such as neural networks. Furthermore, the data set should be expanded to make the model more reliable.

**References**

1. D. Fissore, Freeze Drying of Pharmaceuticals, Encyclopedia of Pharmaceutical Science and Technology, New York, 2013, pp. 1723-1737.
2. H. Grohganz, D. Gildemyn, E. Skibsted, J. M. Flink, J. Rantanen, Towards a robust water content determination of freeze-dried samples by near-infrared spectroscopy, Analytica Chimica Acta, Elsevier, 2010, pp. 34-40.
3. T. De Beer, M. Wiggenhorn, R. Veillon, Importance of using complementary process analyzers for the process monitoring, analysis and understanding of freeze drying, Analytica Chimica Acta, Elsevier, 2009, pp. 7639-7649.
4. T. De Beer, A. Burggraeve, M. Fonteyne, S. Saerens, C. Vervaet, Near infrared and Raman spectroscopy for the in-process monitoring of pharmaceutical production processes, International Journal of Pharmaceutics, Elsevier, 2011, pp. 32-47.
5. S. Bobba, D. Fissore, N. Zinfollino, Evaluation of the Robustness of A Novel NIR-based Technique to Measure the Residual Moisture in Freeze-dried Products, Journal of Pharmaceutical Sciences, In press.
6. A. C. Muller, S. Guido, Introduction to Machine Learning with Python, O’Reilly, 2017.
7. A. Geron, Hands-on Machine Learning with Scikit-Learn, Keras and TensorFlow, O’Reilly, 2019.