6th International Food Operations & Processing Simulation Workshop
17th International Multidisciplinary Modeling & Simulation Multiconference

ISSN 2724-0355 ISBN 978-88-85741-54-6 © 2020 The Authors.
DOI: 10.46354/i3m.2020.foodops.001

Data-Based Herb Contamination Prediction and Harvest Recommendation

Stefan Anlauf[1,2,*], Andreas Haghofer[1,2], Karl Dirnberger[3] and Stephan Winkler[1,2,4]

[1]FFoQSI GmbH, Technopark 1C, 3430 Tulln, Austria
[2]University of Applied Sciences Upper Austria, Bioinformatics, Softwarepark 11, 4232 Hagenberg, Austria
[3]Österreichische Bergkräutergenossenschaft, Thierberg 1, 4192 Hirschbach, Austria
[4]Johannes Keppler Universität, Computer Science, Altenberger Straße 69, 4040 Linz, Austria

[*]Corresponding author. Email address: stefan.anlauf@fh-hagenberg.at

Abstract

The quality of freshly harvested herbs is heavily influenced by multiple factors, namely weather conditions, harvesting, transport, drying, storage, and many more. Our main goal here is to identify models that are able to predict spore contaminations on different types of herbs on the basis of these factors as well as to find optimal processing parameters, which shall lead to lower contaminations of herbs as well as lower costs for contamination prevention represents.

The here presented workflow utilizes two different approaches, which in combination shall lead to a reliable contamination prediction and prevention mechanism. For the prediction part we learn ensembles of machine learning models using the processing parameters as features to predict the risk for spore contamination a priori of labor analysis data. Using tree-based modelling algorithms we already achieved a spore contamination prediction accuracy of 86.21% for the herb nettle. In Addition to that, we use descriptive statistics to provide information on the relevant parameters which could be responsible for the occurred contamination. Here we already achieve a p-value smaller than 0.01 for a few processing parameters.

In the future we want to expand this workflow by improving the modelling process using different modelling algorithms. Additionally, we are working on an online life system, which combine these two methods, to not only present a farmer the information whether a contamination is probably, but also provide him the information which processing parameters lead to a contamination and how they should be affected to lower the risk.

Keywords: data preprocessing; applied statistics; contamination classification; machine learning

1. Introduction and Overview

Even if the weather is for sure one of the most critical factors for the quality of nearly every raw material type in the agriculture sector (Sivakumar, 2007), there are many more variables influencing the quality during the whole production line. In this project, we focus mainly on the data-based identification of crucial factors that influence the quality of herbs processing, which reaches from the planting process to the final harvesting and drying procedures.

Due to the huge number of factors which could lead to a contamination of the final product, it is nearly impossible to do an appropriate analysis without using data science methods. Getting knowledge about the relevance of agriculture factors is especially useful for organic farming facilities, which cannot use the same preventive methods as conventional farmers to inhibit contaminations. If a contamination occurs, this often leads to very high costs for decontamination for the farmers, which reduces their profit. Thus, the goal is to avoid contamination and therefore reduce decontamination costs.

Our methodology uses a combination of machine learning classifiers (Kotsiantis, 2007) and applied statistics to provide the farmers information about possible contaminations, but also offer recommendations for the right treatment to prevent this contamination.

The here presented workflow detects relations between the documented agricultural parameters and the documented contaminations. Using models trained using machine learning, it is possible to predict the risk of contamination with spores such as yeast or mold. Additionally, we use statistical analysis to identify those processing parameters which are most likely to be responsible for spore contamination.

2. Data

As shown in Figure 1, the actual data used for preprocessing and later on for machine learning and statistical analysis is compiled of various information sources storing data about the processing of herbs on the one hand and the laboratory information about contamination of batches on the other hand.

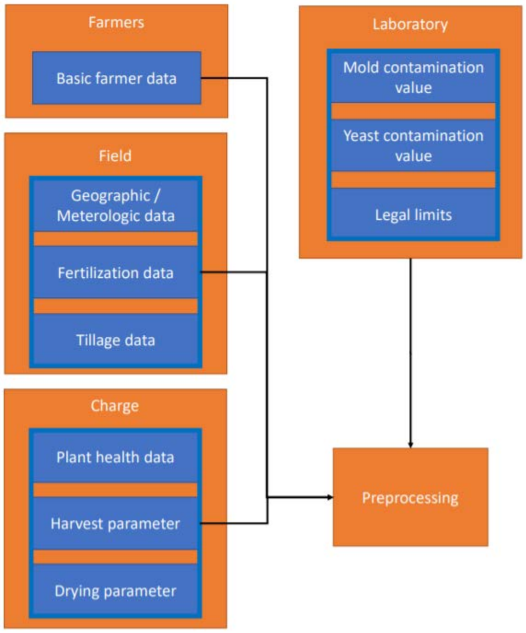

Figure 1: Structure and source of the two types of data (processing data; laboratory data) and how they are combined for the preprocessing step

2.1. Processing Data

To be able to evaluate significant parameters which have an impact on the final product quality, it is necessary to collect all data from the point where the seed is planted on the field, until the final product is ready to be packed.

We have developed a survey to collect data from 60 different farmers over two years including the following categories of parameter data:

- basic data of the farmer and the field (sea level, field size)
- meteorological data of the field
- fertilization information
- tillage information
- visual impression of the condition of a raw material sample
- post-harvest treatment parameters
- harvesting parameters
- drying parameters
- weather on harvest day

Each of these categories is documented for each harvested batch for which samples are analyzed by the laboratory to determine the spore load. The combined data of all these categories results in 178 different features including nominal, ordinal and metric data. These data are collected for the following raw materials:

- nettle
- peppermint
- apple mint
- oregano
- dandelion
- lemon balm

2.2. Laboratory Analytics Data

As ground truth information, whether there is any measurable contamination within the harvested crops, our system relies on the laboratory analysis of harvest samples. This analysis contains the exact number of spores or bacteria measured within the sample and the information if this value is within the legal limits. At the current state of our project, we use the analytics for spore contaminations:

- mold
- yeast

3. Methods

Each of the data samples submitted by the farmer first undergoes some preprocessing steps. Prediction models for contamination are trained using machine learning, and later these models can be applied on new samples stating whether a contamination is to be expected or not. Additionally, statistical analysis is used to recommend processing parameters by analyzing the values of all given parameters in samples without contamination versus the corresponding values in samples for which a contamination was detected.

3.1. Data Preprocessing

As in nearly any data science project, especially those were the data is collected by a survey, we must deal with data quality problems such as missing-values and noisy data.

These problems had to be solved twice because the preprocessing of the statistics part was not the same as the one for the data used to train machine learning-based models. Based on mathematical operations, most machine learning algorithms and statistical operations cannot cope with verbal information as features. Therefore, it is necessary to map each character value of a feature column to unique numeric values.

Due to the heterogeneity of the data origins, it is necessary to split the dataset into subsets for each specific raw material type. This splitting process results in one dataset for each raw material type. Another vital information gathered during the first stages of this project with these data was the fact that it is not possible to train one model which can predict all different contaminations correctly. The used approach to solve this problem was to train one model for each combination of contamination and raw material type.

Even if all values of the harvest data if required to be entered for every batch, it is often not possible for the farmers due to various circumstances. This lack of information leads to missing values within the final dataset (Streiner, 2002). If there were plenty of data, it would be no problem to remove each data entry, which does not provide every required value, but this is not possible in our case.

The most frequently used strategy for dealing with missing values is to use the information of the other data entries to replace the missing values with the mean or median of the other entries (Acock, 2005). We use this strategy for all numeric values. For nominal and ordinal datatypes, we use the median. Both strategies only work correctly if the amount of missing values for the current feature is low enough to inhibit the problem of bias and noise due to the inserted mean or median value (Acock, 2005). If a column has more than one-third missing values, the whole column was removed from the dataset.

For an improvement of the final model quality, it is common to apply feature selection methods on the dataset. For this step, we use the correlation between the features to select only the ones with a low correlation with the rest of the features. (Hall, 1999). Applying this idea results in a dataset that should only contain features providing information for the target prediction, which could not be provided by the other features. This feature selection also includes a filter for features with the same value for each entry, which does not provide any useful information.

For the statistical analysis, the preprocessing the data is grouped by every type of raw material. In addition to that, the data are split into contaminated and non-contaminated batches. This data separation requires on the fly filtering of the whole database and therefore results in a temporary dataset for each analysis.

3.2. Learning Contamination Prediction Models

The preprocessed data is used to train different machine learning classifiers that classify new batches as contaminated or not contaminated. As described in Chapter 4.1 theses classifiers are later used for the spore contamination prediction of new batches. Figure 2 shows the pipeline for learning models using machine learning.

To use as much data as possible for the training and to provide a reliable validation methodology, leave one out cross-validation represents a valid alternative to the more traditional training and test split method (Webb, 2011). I.e., each sample is in turn used as test sample and all others as training samples. Thus, we retrieve a high number of models and a good estimation of the achievable accuracy.

This cross-validation modeling step is executed for each type of raw material and each contamination. This modeling process results in contamination prediction models for each raw material type.

In our system, we filter the models and use only those that are better than baseline accuracy. This threshold (baseline) is calculated by calculating the ratio of the most frequent target value (contaminated or not contaminated) in all entries. Our filter requires the models to be at least ten percent better than this baseline value.

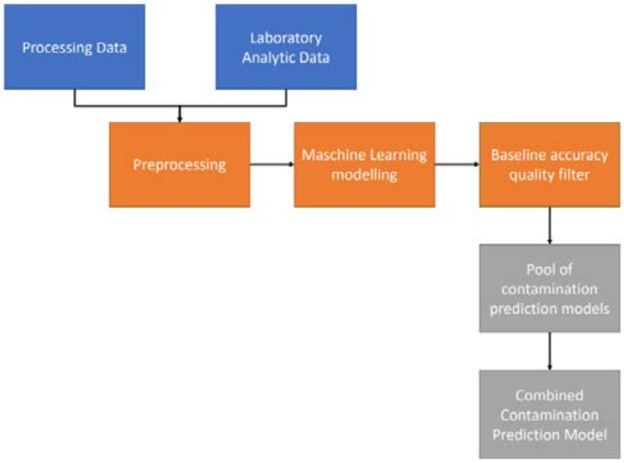
Processing Data
Laboratory Analytic Data
Preprocessing
Maschine Learning modelling
Baseline accuracy quality filter
Pool of contamination prediction models
Combined Contamination Prediction Model

Figure 2: The exact processing-workflow for training new machine learning models used for contamination prediction including the quality filter using the data described in Figure 1.

Each model that passes this quality criterion becomes a part of the model pool used for the contamination prediction. The generated pool of machine learning models in the current state of our project is not able to predict the chance for contamination for every combination of raw material type and contamination. Therefore, we provide a retrain potential into our pipeline, which results in new models replacing the old models, every time new laboratory data is received. Due to the need for the laboratory results for the modeling process as target values, each new data entry of the dataset could only be used for the modeling when the laboratory results arrived.

Our system not only relies on one type of model for one combination of raw material type and contamination. Instead, we use two different tree-based modelling classifiers algorithms, random forest (Breiman, 2001) and gradient boosting (Chen, 2016). For the implementation of these two different algorithms we use the open source programming language python. More specifically we use the package Scikit-learn, which already contains the implementation of random forest and xgboost with the standard parameter settings (Pedregosa, 2011).

Not only the best performing model of the available ones can be used for contamination prediction, but it is also possible to treat all available models or a selection of them as an ensemble (Dietterich, 1995) and use the mean of all model predictions as final result.

3.3. Identification of Crucial Parameters by Statistical Analysis

We use hypothesis tests to extract the information, which of the processing parameters are responsible for spore contaminations. We analyze the differences of the distributions of the parameter values in contaminated samples and those in not contaminated samples. Depending on the data type of each processing parameter, the selected statistical tests calculate the significance of these differences between the non-contaminated and contaminated data entries for each raw material type and contamination. Using a significance level of 5%, only the processing parameters below this level are selected and therefore regarded significant influence factors for the contamination. Figure 3 shows the here applied analysis pipeline. The hypotheses tests used in this project are listed in Table 1.

Table 1: Hypothesis test methods for the different data types

Nominal Data	Pearson's Chi Squared Test (Pearson 1992)
Ordinal Data	Mann–Whitney U Test (Gooch, 2011)
Metric Data	Two Sample F-Test for testing variance homogeneity (Schumacker, 2013)
Metric Data (variance homogenous)	Mann–Whitney U Test (Gooch, 2011)
Metric Data (variance heterogenous)	Two Sample T-Test (Schumacker, 2013)

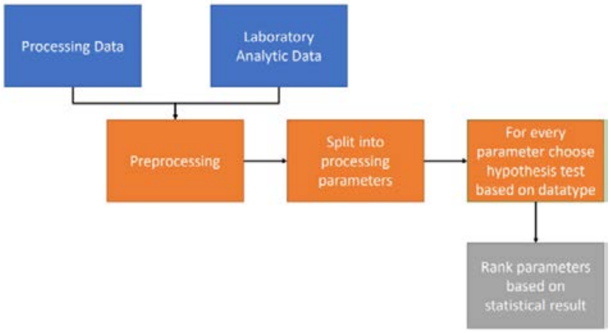
Processing Data
Laboratory Analytic Data
Preprocessing
Split into processing parameters
For every parameter choose hypothesis test based on datatype
Rank parameters based on statistical result

Figure 3: The exact processing-workflow of using hypothesis tests to rank the processing parameters based on their relevance using the data described in Figure 1.

After the preprocessing step, the treatment data of one type of raw material is split into all single processing parameters. So, all parameters can be analyzed on their own. In addition to the parameter the user also can see to which category this parameter belongs. The parameters are ranked, based on the resulted p-value of the hypothesis tests. So, the user gets a list of all relevant parameters, starting with the most relevant one. The relevance is calculated by subtracting the p-value from 100 and multiply this result by 100. The biggest difference between this analysis and the recommendation system is, that this analysis only uses old data. The analysis always calculates on the fly using the current data pool of treatment data from the database, from which the laboratory results are available. The same procedure also is used for the recommendation after a positive prediction.

Table 2: Listed results of different raw material contamination combinations and the different machine learning approaches.

Contamination	Material	Model	Number of samples	TP	FP	TN	FN	ACC
mold	nettle	RF	58	58.5 %	12 %	27.5 %	2 %	86.21 %
yeast	peppermint	RF	64	56 %	8 %	33 %	3 %	81.25 %
yeast	peppermint	xgboost	64	50 %	9.5 %	31 %	9.5 %	85.94 %

4. Results and Discussion

Both methods, the statistical analysis and the machine learning prediction, both use always the data of one raw material type and contamination combination. So currently it is impossible to present results for all these combinations, a few combinations of different types of raw material with the most frequent contaminations yeast and mold are possible.

4.1. Prediction of spore contaminations yeast and mold using machine learning classifiers

In the following chapter, we present two different contamination classification results, one for each type of contamination with different raw material types. The combined results of both combinations and different machine learning approaches are presented in Table 2.

4.1.1. Prediction of mold contamination in nettle using random forest classification

Trained on 58 samples of nettle batches, this random forest model could achieve a test accuracy of 86.21 %. To achieve that our implementation of random forest uses a lot of the standard parameters for random forests. The number of features used for every tree is the square root of the whole number of features. In addition to that, we set the number of trees to 51.

As seen in Table 3, this result is quite good because of the small false negative and false positive rate. Especially, the very small false negative rate is noteworthy. The results also show that the prediction of mold in nettle achieve a bit higher accuracy than the prediction of yeas in peppermint (Chapter 4.1.2).

4.1.2. Prediction of yeast contamination in peppermint using random forest and gradient boosting classification

Unlike the example for mold prediction, this example of yeast prediction on peppermint data allowed the use of a random forest model in combination with a xgboost model, trained on 64 samples. The random forest model uses the same parameter settings as in Chapter 4.1.1, but the xgboost model uses a bit different ones. We use the same number of features per tree, but instead of 51 trees the number is increased to 201 trees. The number of samples used for every tree is set to 30% of all samples.

Table 2 shows the prediction results of the two different tree-based approaches. With a test accuracy of 85,94 % the xgboost model is slightly better than the random forest model with 81,25 %. The prediction of the xgboost model is more reliable. In this case the previously mentioned ensemble of the two models would underperform with a combined test accuracy of 78,13 % compared to the xgboost model alone with 85,94 %. So, the best model to use is the xgboost model alone, not only because of the high test accuracy, also again because of the very small false negative rate with 3 %.

4.2. Nettle processing parameter analysis for yeast contamination

For the exact combination of the raw material nettle and the contamination yeast we achieved a very good parameter analysis, because of the high amount of data entries. This analysis results in a list of processing parameters that are relevant for a yeast contamination and can reduce the chance of yeast contamination by adapting those parameters.

Although a p-value is calculated for all processing parameters, in the following, we just present the most relevant and few not relevant processing parameter results. The most relevant area is the feature category “field and crops”, more specific the part of the processing parameters that happen before or while the herb is growing.

Table 3: List of a few processing parameters with the used hypothesis test, ranked based on their results (p-value).

Parameter	Category	Hypothesis Test	P-Value
weed	field and crops	Pearson’s Chi Squared	0.001
fertilization in fall	field and crops	Two Sample T-Test	0.002
preheated	drying process	Pearson’s Chi Squared	0.004
height of the plant stick	field and crops	Two Sample T-Test	0.01
cutting device freshly sharped	harvesting	Pearson’s Chi Squared	0.11
sea level	basic data	Two Sample T-Test	0.18

As shown in Table 3, the most relevant parameter is the decision whether you leave the weed on the field or if you remove the weed. The performed Pearson’s Chi Squared Test results in a p-value of 0.001. 39 out of 54 from all the not contaminated batches leaves the weed on the field, while 8 out of 12 from the bad batches export it.

Another very relevant “field and crops” parameter is the amount of fertilization in fall. With this parameter values the Two Sample T-Test calculated a p-value of 0.002. On average, the good batches are fertilized more (26 t) than the bad batches (23.3 t).

The third relevant parameter of this category is the height of the plant stick from where the plant starts growing. The p-value again calculated with the T-Test is 0.01. So, the relevance is slightly lower, but still significant at the level of 0.05. Here the average of all good batches with 6.25 cm height is a bit higher than the average of the bad ones with 5 cm.

The parameter, if you preheat your dryer before starting drying has a calculated p-value of 0.004 using the Pearson’s Chi Squared Test. But this parameter belongs to a completely different category, the drying process. 43 out of 56 of all not contaminated batches have preheated their dryer, which is a percentage of around 77 %. Considering the contaminated batches only 8 out of 13 (62%) have preheated their dryer.

As mentioned above, there are also a lot of processing parameters that are not significant for the yeast contamination. A small example of those parameters is presented.

A slightly not significant parameter is the sharpness of the used cutting device. The performed Pearson’s Chi Squared Test results in a p-value of 0.11, so the parameter if your cutting device is freshly sharped, is not significant for a yeast contamination.

Another not relevant parameter is the sea level of the field. With a p-value of 0.18 resulting from the T-Test the sea level is not significant. Here the average of all good batches with 651m height is just a bit lower than the average of the bad ones with 709 m, but not significantly.

5. Conclusion and Outlook

In the current state it is already possible to organize the data, link the treatment data with the laboratory results and get some accurate prediction and recommendation results.

Combining the results from the machine learning prediction and the statistical analysis, this is the point where in the future the farmers can profit. Now for the combination of nettle and yeast he or she can have accurate prediction about his current processing parameters and furthermore he or she gets a list which parameter are the most relevant for the yeast prediction. So, editing those parameters will lead to a less probably contamination and in the best case safe the decontamination costs. In the future this whole workflow should end up in an online system, which the farmer can use themselves and see their results and recommendations.

Furthermore, the results also can improve by adding more machine learning methods, such as for example neural networks and symbolic regression. For yeast prediction already two different methods are used. In the future the number of available prediction models should increase. Also, the amount of data, that can be used for machine learning or statistical analysis, increases over time. So, the number of models and the accuracy will increase, and the relevance of the most relevant processing parameters increases if the still stay relevant with new data.

This methodology already provides the potential to be used for any agricultural data and therefore has the potential to impact an overall reduction of contaminated batches actively. In further version of this pipeline we also want to include transfer learning to improve the overall quality and the amount of our model pool.

References

Sivakumar, Mannava V K and Motha, Raymond P. (2007). Managing Weather and Climate Risks in Agriculture. *ISBN 978- 3540727446.*

Kotsiantis, S. B. (2007). Supervised machine learning: A review of classification techniques. *Informatica (Ljubljana)*

Streiner, David L. (2002). The case of the missing data: Methods of dealing with dropouts and other research vagaries. *Canadian Journal of Psychiatry*

Acock, Alan C. (2005). Working with missing values. *Journal of Marriage and Family pp. 1012-1028*

Hall, Ma (1999). Correlation-based feature selection for machine learning. *Diss. The University of Waikato. ISBN 978-0874216561*

Dietterich, Thomas G. (2000) Ensemble methods in machine learning. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics). ISBN 354-0677046*

Webb, Geoffrey I. and Sammut, Claude and Perlich, Claudia and Horvath, Tamas and Wrobel, Stefan and Korb, Kevin B. and Noble, William Stafford and Leslie, Christina and Lagoudakis, Michail G. and Quadrianto, Novi and Buntine, Wray L. and Quadrianto, Novi and Buntine, Wray L. and Getoor, Lise and Namata, Galileo and Getoor, Lise and Han, Xin Jin, Jiawei and Ting, Jo-Anne and Vijayakumar, Sethu and Schaal, Stefan and Raedt, Luc De. (2011). Leave-One-Out Cross-Validation. *Encyclopedia of Machine Learning. ISBN 978-0387307688 pp. 600 - 601*

Chen, Tianqi and Guestrin, Carlos (2016) XGBoost: A Scalable Tree Boosting System. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*

Breiman, Leo (2001). Random forests. *Machine Learning pp. 5-32.*

Pearson, Karl (1992). On the Criterion that a Given System of Deviations from the Probable in the Case of a Correlated System of Variables is Such that it Can be Reasonably Supposed to have Arisen from Random Sampling. *Breakthroughs in Statistics: Methodology and Distribution pp. 11--28*

Gooch, Jan W. (2011). Mann-Whitney U Test. *Encyclopedic Dictionary of Polymers*

Schumacker, Randall and Tomek, Sara (2013). F-Test. *Understanding Statistics Using R pp. 187 -207*

Pedregosa, Fabian and Varoquaux, Gael and Gramfort, Alexandre and Michel, Vincent and Thirion, Bertrand and Grisel, Olivier and Blondel, Mathieu and Prettenhofer, Peter and Weiss, Ron and Dubourg, Vincent and Vanderplas, Jake and Passos, Alexandre and Cournapeau, David and Brucher, Matthieu and Perrot, Matthieu and Duchesnay, Edouard (2011). Scikit-Learn: Machine Learning in Python. *J. Mach. Learn. Res. pp. 2825 - 2830*