MASTER

Optimization of agricultural raw material allocation
a case study in potato product manufacturing

Willems, N.

Award date:
2019

Manufacturing Systems Engineering
Dynamics & Control
Department of Mechanical Engineering
Master Thesis - DC 2019.059

Optimization of Agricultural Raw Material Allocation: A Case Study in Potato Product Manufacturing

By:
N. Willems

ID Number:
0851016

Supervisors Royal HaskoningDHV:
E.A. Putri
P.L.J.C. Ramakers

Supervisor TU Eindhoven:
prof.dr.ir. I.J.B.F. Adan

Supervisor Lamb Weston / Meijer:
H. van der Mars

Eindhoven, June 4, 2019

Optimization of Agricultural Raw Material Allocation: A Case Study in Potato Product Manufacturing

Nick Willems[a,b,*]

[a] *University of Technology Eindhoven, Department of Mechanical Engineering, Eindhoven, The Netherlands*
[b] *Royal HaskoningDHV, Project Management & Consultancy Multinationals, Eindhoven, The Netherlands*

Abstract

Raw material to end product allocation can lead to big processing improvements when executed optimally. In this paper a case study is presented within potato product manufacturing to find optimal allocation of potatoes, end products and manufacturing lines. The allocation problem is formulated as a linear program which aims to minimize raw material input. This program requires as input the factory recovery of every combination of material, product and manufacturing line. Hence, to solve this program, predictions on factory recovery need to be made. For this purpose a nearest neighbour interpolation based algorithm and a random forest based algorithm are developed. Those techniques are applied to a data set of historical production orders. The proposed approach of linear programming and recovery prediction indicates potential savings in raw material use and purchasing costs. The approach is generally applicable and could lead to possible savings in similar types of processing industries as well.

Keywords: raw material allocation, linear programming, data mining, factory recovery, crop processing

1. Introduction

Although nowadays farmers' working methods and prediction tools are better developed than before, uncertainty in harvest during the year will always be there. This

[*] Corresponding author
Email address: nickwillems38@hotmail.com (Nick Willems)

expresses itself in 1) differences in characteristics and quality of the harvested material and 2) differences in the size of the harvest. In industry, variability in raw material characteristics must be taken care of to generate the needed amount of end products one way or another. Differences in characteristics lead to the fact that some material is better suited for a particular end product than other materials. These differences make allocation from crops and fruits to product a difficult process, since wrong allocation leads to a lot of processing losses. Obviously this should be prevented for environmental, financial and social reasons. According to the FAO, the Food and Agriculture Organization of the United Nations, approximately 800 million tonnes of root and tubers crops are initially produced each year [1]. Ten percent of the initial production of root and tuber crops got wasted during processing. For companies which produce products out of agricultural raw material the processing steps are schematically displayed in Figure 1. The general idea is that even when the process is considered as a "black box", relations between raw material, product and processing efficiency exist. It is necessary to find out which characteristics of these three factors lead to particular waste amounts. Variety in raw material properties is often reflected by differences in sizes and weights and variety of the material. The manufacturing process can often be performed on multiple manufacturing lines, making allocation even more complicated. Also the product may be divided into homogeneous groups of end product based on specific characteristics.

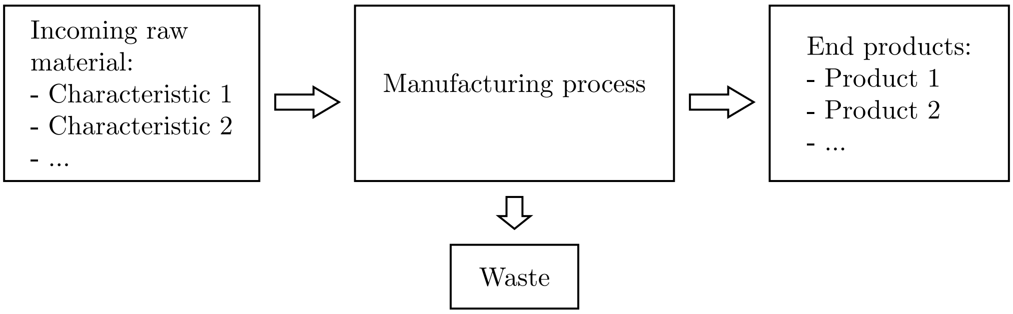


Figure 1: Simplified visualisation of the manufacturing process for food processing companies.

According to Henningsson et al. [2], for the food and drink industry, waste minimization savings in raw materials by either changes in technology or procedures carry the greatest potential for financial savings. In particular, the savings achieved through low-cost procedural changes are remarkable.

Also from more recent work it appears that there is a call for food waste prevention initiatives. Papargyropoulou et al. [3] state that prevention, through minimization of food surplus and avoidable food waste, is the most attractive option as a first step towards a more sustainable resolution.

In Zhong et al. [4] data-driven intelligent manufacturing models are opted to improve manufacturing efficiency, but also as a driver of greater agility and deeper integration with other parties such as logistics and supply-chain management entities. Data-driven models are able to make full use of historic or real-time data for system diagnosis or prognosis, based on information or knowledge integration, data mining and data analysis.

In this paper optimal allocation of raw material is opted to improve yield. A case study at a potato product manufacturer is performed to reallocate their raw material. The allocation of raw material is based on **Factory Recovery**. Factory recovery is one of the variables captured to address losses during the manufacturing process. It is the percentage that indicates how much of the raw material fed into the manufacturing line is turned into useful product in terms of weight. When the factory recovery is high, the amount of losses during the process is low. To illustrate, if 1,000 tonnes of raw material is used to generate 600 tonnes of product, the factory recovery is 60%. To reallocate resources and prevent losses, the factory recoveries of every possible combination of potatoes, products and manufacturing lines needs to be predicted with high accuracy. For this prediction two data-driven techniques are elaborated, to wit, nearest neighbour interpolation and random forests for regression.

2. Problem Description

In general, companies generate multiple stock keeping units (SKUs) from varying raw material (types) on one or multiple manufacturing lines. At the company considered for this research, raw material consists of potatoes, which distinguish themselves based on variety, length and underwater weight. These potatoes are processed to end products, mainly different types of french fries, by executing a production order on one of the nine available manufacturing lines of this manufacturer. The end products distinguish themselves based on cutting style, grading level (e.g. regular, premium) and skin presence. For the purpose of production an annual recurring process is performed which is based on the sales forecast. This process is visualized in Figure 2.

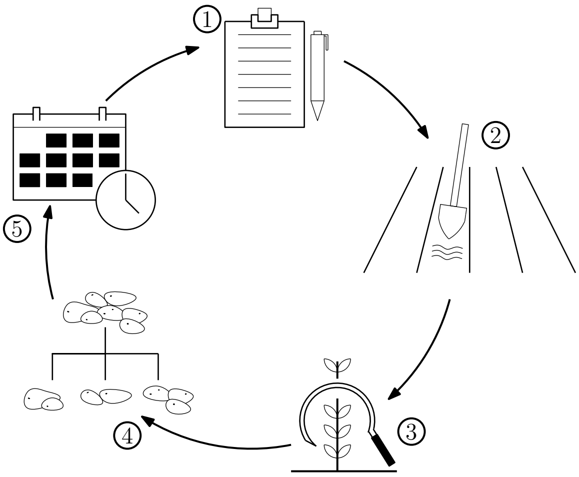


Figure 2: The major steps in the annual recurring process cycle of the potato product manufacturer with 1) contracting, 2) planting, 3) harvesting, 4) allocation and 5) scheduling.

The process starts by contracting suppliers for the new harvest year by means of the sales forecast combined with suppliers' historical performance. The contracting phase takes place in January and February. After that, in March, the planting of the tubers is executed by the farmers and then the potatoes are harvested in the period from July to October. From this moment the allocation process starts, since the availability of raw material and the sales forecasts are known. Scheduling the execution of the allocated production orders is the final step. The focus of this paper is on the allocation process of the potato product manufacturer; this process is schematically displayed in Figure 3.

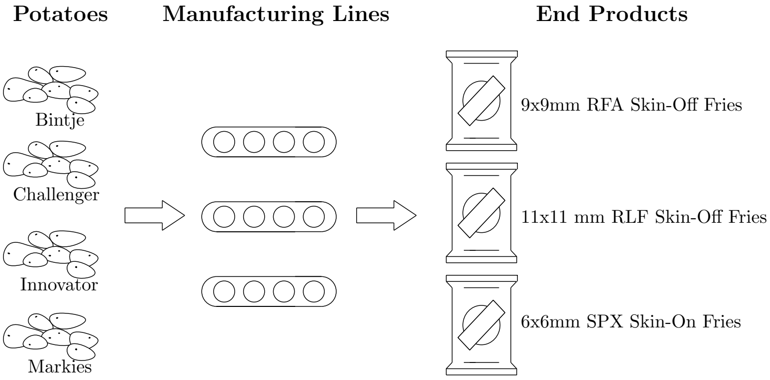


Figure 3: Allocation process for the potato product manufacturer.

Allocation is considered to be a wasteful process when it is executed non optimally. To test this assumption and to come up with a way to prevent these losses as much as possible, a case study is conducted. The goal is to find out whether the production orders of the potato product manufacturer could have been made from the same stock using less potatoes. From July 2016 to July 2018 a set of 3,006 executed production orders is constructed from which it is known how many potatoes were used to produce a certain amount of product. It will be investigated whether it is possible to rearrange the resources to obtain the same amount of product with lower potato input. Since 50 potato varieties ($N = 50$), 183 product types ($M = 183$) and 9 manufacturing lines ($L = 9$) occur in these orders, the factory recovery of 82,350 combinations needs to be determined. It is necessary to find out which data and techniques are available for accurate prediction of recovery.

3. Data Preparation

Zhong et al. [4] proposed to use data-driven models to improve manufacturing efficiency. A data-driven approach has also been chosen here. The entire process of data preparation, data analysis and model creation for this case study is performed following the cross-industry standard process for data mining (CRISP-DM). The CRISP-DM methodology provides a structured approach to planning and executing a data mining project [5]. The life cycle of a data mining project can be divided into six major phases, displayed in Figure 4.

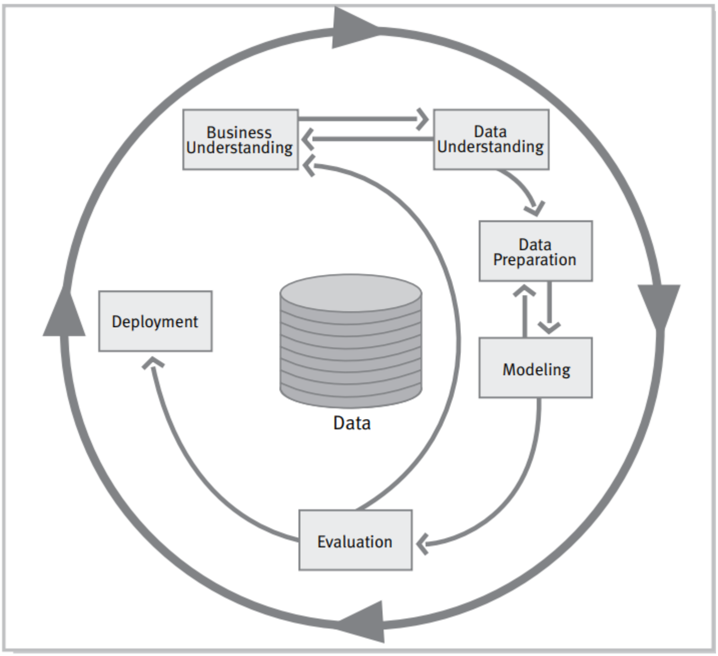


Figure 4: Phases of the CRISP-DM reference model [5].

The initial phase, business understanding, focuses on understanding the project objectives and requirements from a business perspective. As elaborated, the project objective is to minimize losses during the manufacturing process by means of better raw material allocation. To do so, it is necessary to be able to predict every combination of potatoes, products and manufacturing lines and thus make prediction possible. The second phase is data understanding, which starts with initial data collection. Data preparation consists of constructing the final data set, in other words the data set that will be fed into the modelling tools. In the modelling phase, various modelling techniques are selected and applied, and their parameters are calibrated to optimal values. During the evaluation phase it is evaluated whether the model created properly achieves the objective; is it able to predict every combination with high accuracy? Finally, the results of the generated model needs to be made practicable during the deployment phase. The knowledge obtained, in this case the factory recoveries, should become accessible and should be processed for further use, to wit, for reallocation of the potatoes. The data analysis and data preparation steps are elaborated in the remainder of this section. Modelling and evaluation are described in the coming section. Thereafter, the deployment phase is covered which describes the use of the recoveries for reallocation of the potatoes.

Data mining techniques, applied during the modelling phase, are usually executed using a set of data typically looking as shown in Table 1. Also for regression, one of the common tasks within data mining, a set like this is favourable.

Table 1: Typical data set for creation of data mining models.

Case	**Feature 1**	**Feature 2**	**...**	**Feature** n	**Target value**
1	xxx	x		xx	y
2	xxx	x		xx	y
3	xxx	x		xx	y
...					

However, in our case the data needs a lot of preprocessing before it looks like the data set in Table 1. The set, with factory recovery as target value, needs to be constructed from separated files. Those files and the links between them are displayed in a Data Structure Diagram in Figure 5. As stated before, this data contains production orders executed from July 2016 till July 2018 which concerned fries and specialities, excluding potato wedge products. It contains a lot of parameters, but it may be possible that not all parameters displayed are necessary for an accurate prediction. Therefore, first analysis is conducted to identify data quality problems and to discover first insights into the data.

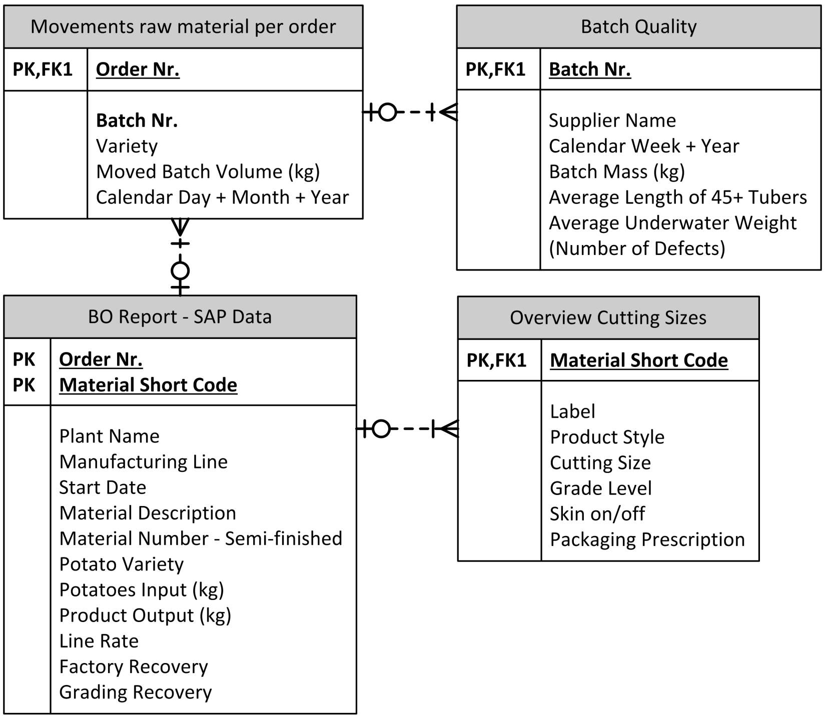


Figure 5: Data structure diagram of the available data sheets with possible predictor values.

This analysis led to the insight that a lot of numerical data about potato characteristics was missing. Besides that, the quality of the numerical data that was available was questionable due to the batch sampling process. The recording of the length and underwater weight of incoming potato batches is a rather deficient process and traceability is limited. For this reason it was decided to exclude the numerical data from potato characteristics in this research. Thereafter, outliers have been removed if they differed more than 1.5 times the standard deviation of the average of the group they were in. A group consists of orders which have equal categorical values for all parameters. What remained was a single data set with the characteristics displayed in Table 2. A sample of this set is displayed in Table 9 in Appendix A.

Table 2: Characteristics of the final data set.

Number of records	**Number of groups**	**Average Factory Recovery**	**Std. Dev. Factory Recovery**
3006	514	61.4%	7.8%

In this set 514 unique groups are found, which means that 514 unique combinations of categorical variables are found. The categorical predictors that are used in this paper are 1) potato variety, 2) cutting style of the product, 3) skin presence, 4) grade level of the product and 5) manufacturing line to produce on. Examples of values that these variables can adopt are displayed in Table 3.

Table 3: Categorical predictors with example values they can adopt.

Predictor variable	**Classes**	**Example of class values**
Potato variety	50	Bintje, Innovator, Markies, Fontane
Cutting style	39	6x6mm, 9x9mm, Twister Fries, 9x18mm
Skin presence	2	Yes, No
Grade level	11	RFA, RFW, SFW, SFY, SPX, SXL
Manufacturing Line	9	0020-1, 0020-2, 0030-1, 0040-1, 0050-1

4. Model

The process of allocation of agricultural raw material to end products to be generated can be written as a Linear Program (LP). The objective of this LP is minimizing potato input, while still being able to generate the desired end product amounts. Constraints should be imposed to make the LP realistic and complete for the described scenario.

The first of these constraints follows from the fact that it is not possible to feed more potatoes of a specific type into one of the manufacturing lines than there are potatoes of that type available. Thus, the total fraction of a material type fed into the lines should be equal to or smaller than one.

The second constraint arises from the fact that certain product amounts should be generated. The amounts correspond to either expected sales, ordered sales or historical sales. In this study the demand is based on historical generated product amounts. The product that will be generated is equal to the amount of potatoes of a type fed into the manufacturing lines times the recovery of that combination.

In the third place, a constraint with respect to manufacturing line capacity is needed. Every line can process a certain amount of potatoes. It is chosen to limit capacity based on a maximal amount of product that can be generated on each of the lines, since this is how the manufacturer estimates their capacity. If this capacity constraint is not imposed, it can be possible that the outcome of the problem is to produce all orders on a particular line. From one point of view this would be a useful insight; this line can be considered as the most effective. However, in practice it is impossible to produce all orders on one manufacturing line and pause the others.

Then, some products can be made out of specific varieties only, mainly due to customer demands or product characteristics. This implies that varieties, other than allowed according to the contracts or suited for production, should not be used.

Besides that, some manufacturing lines are not suited for the production of particular products. Necessary machines may not be present in the line or certain settings, e.g. cutting accuracy, can not be met.

Finally, a non-negativity constraint is needed since material input can not be negative. All constraints above lead to an LP which reads as follows;

$$\begin{aligned}
&\text{minimize} && \sum_{i=1}^{N}\sum_{j=1}^{M}\sum_{k=1}^{L} S_i x_{i,j,k} && && (1)\\
&\text{subject to} && \sum_{j=1}^{M}\sum_{k=1}^{L} x_{i,j,k} \leq 1, && i = 1,2,...,N && (2)\\
& && \sum_{i=1}^{N}\sum_{k=1}^{L} S_i R_{i,j,k} x_{i,j,k} \geq D_j, && j = 1,2,...,M && (3)\\
& && \sum_{i=1}^{N}\sum_{j=1}^{M} S_i R_{i,j,k} x_{i,j,k} \leq C_k, && k = 1,2,...,L && (4)\\
& && \sum_{i\in\mathbf{I}_j}^{N}\sum_{k=1}^{L} x_{i,j,k} = 0, && j = 1,2,...,M && (5)\\
& && \sum_{k\in\mathbf{K}_j}^{L}\sum_{i=1}^{N} x_{i,j,k} = 0, && j = 1,2,...,M. && (6)\\
& && x_{i,j,k} \geq 0, && \text{for all } i,j,k. && (7)
\end{aligned}$$

The goal is to find decision variables $x_{i,j,k}$, the fraction of material i used for the production of product j on manufacturing line k, which minimize potato input and do not violate the constraints (1). In this formulation S_i is the available amount of potatoes per variety. The constraint on potato input is denoted in (2). The amount of product that needs to be generated is denoted by D_j. This is based on the raw material input and the factory recovery $R_{i,j,k}$ (3). Every manufacturing line has a certain capacity C_k, which is set in the form of a maximal amount of product that can be made in a time period (4). Then, the constraint on allowed potato type per product is introduced. The set $\mathbf{I}_j$ contains the varieties that are not allowed per product type j. The sum of these fractions should be zero (5), which implies, together with the non-negativity constraint (7), that the corresponding fractions should be zero. In a similar manner the manufacturing lines that are not allowed to produce on are suppressed. The set $\mathbf{K}_j$ contains the manufacturing lines on which product j can not be produced (6).

For the case study, the program above will be solved using the historical amount of potatoes that was used for production as S_i and the amounts of products generated will be used as D_j. To date, factory recovery is not known for all combinations of material, product and manufacturing line and although it is known for some combinations variation occurs. Therefore accurate predictions should be made and used for solving the program.

5. Recovery Prediction

As stated in the previous paragraph, the factory recovery for each combination of potatoes, product and manufacturing line needs to be approximated with an accuracy as high as possible. If combinations are executed before, an indication of the recovery can be traced relatively simple. To predict the factory recovery of combinations that were not made before several data mining techniques are available [6, 7, 8]. Those techniques rely on sufficient historical data, which contains predictor variables and target values (factory recovery). At all times it should be found out which variables are necessary and available for good prediction of recovery. For this research nearest neighbour interpolation and random forests for regression are used. The predictions obtained using nearest neighbour interpolation provide a lot of transparency, since they are based on production orders executed in the past. Besides that, it has proved to be a suitable method for prediction in the research of Jerez et al. [9]. Random forests can handle a lot of different feature types, like binary, categorical and numerical features. Besides that, they are known for their good performance in combination with relatively simple development, which is supported by De'ath & Fabricius [10].

To compare different models and their accuracy, several performance measures are introduced. The first performance indicator is the Mean Absolute Error (MAE), which finds the average difference between the predicted value x_i and the "true" value y_i. MAE is calculated as follows

$$MAE = \frac{\sum_{i=1}^{n} |x_i - y_i|}{n} \tag{8}$$

with n the number of points to be predicted. The second indicator is the Root Means-Square Error (RMSE), which is comparable to the MAE in a way that it also compares the predicted value x_i and the "true" value y_i. RMSE is calculated using

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n}(x_i - y_i)^2}{n}}. \tag{9}$$

The third performance measure used to compare the models is based on the standard deviation of the factory recovery in groups with identical predictor variables. It appears that even if all the input variables of orders are equal, they may not have equal factory recoveries. This can be dedicated to natural variation in the manufacturing process. To illustrate this, an example is used; three historical orders with the same input variables, thus forming a group, had recoveries of 60%, 61% and 63%. If a created model predicts 62% for this combination of input variables, it is not exactly correct for all three orders. However, it can be stated that the prediction is rather accurate and is located within the natural variation. To capture this phenomenon, a third performance measure is introduced. This is the percentage of the total predictions that lie within 1.5 times the standard deviation σ of the average of the recoveries of the group. Thus, for each group a lower and upper bound can be determined and it can be checked whether the predictions are located within these bounds. For the example group the average is 61.3% and the standard deviation σ is 1.53%. The lower bound is then $61.3 - 1.5 * 1.53 = 59.0\%$ and the upper bound is $61.3 + 1.5 * 1.53 = 63.6\%$. The predicted recovery of 62% lies within these bounds and is therefore considered to be accurate enough. This measure has one disadvantage: a group consisting of one order, so a unique production order in the data, has a standard deviation equal to zero. This implies that the actual recovery does not have an upper and lower bound and thus the predicted recovery will never lie within

1.5σ of the average. For this reason the final performance measure is introduced. It is the percentage of predictions that lie within 1.5σ of the average of the actual recoveries, which is the same as the third one, but it excludes groups of size one.

5.1. Nearest Neighbour Interpolation

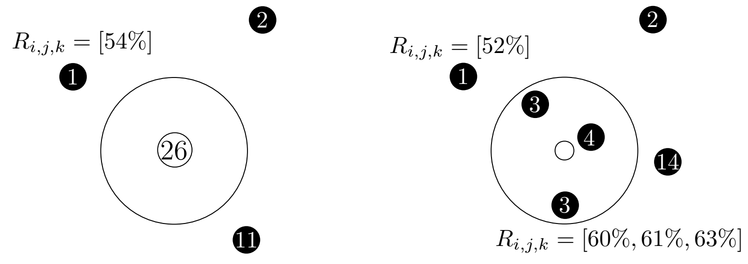
The proposed nearest neighbour interpolation (kNN) algorithm compares a production order from which the factory recovery needs to be predicted with historical orders from the constructed data set. This comparison is based on the five categorical predictor variables ($C = 5$) which should be known for the historical orders as well for the order to be predicted. The extent to which the historical order $X = (X_1, ..., X_C)$ matches the order to be predicted $Y = (Y_1, ..., Y_C)$ can be expressed in a distance measure. This distance measure is based on overlap in categorical predictors, which is expressed as follows

$$D(X,Y) = \sum_{k=1}^{C} \gamma_k d(X_k, Y_k) \tag{10}$$

where

$$d(X_k, Y_k) = \begin{cases} 1, & \text{if } X_k \neq Y_k \\ 0, & \text{if } X_k = Y_k. \end{cases} \tag{11}$$

X_k is the categorical value of predictor k for the order to be predicted and Y_k is the categorical value of predictor k for the historical order. In (10), γ_k is an importance factor. There are five input variables and they can differ drastically in terms of their impact on the response. For this reason the introduction of importance factors is proposed by Joseph & Kang [11]. By parameter tuning and cross-validation on the data set, the optimal importance factors are found to be γ = [2.28, 3.87, 2.81, 2.11, 3.23] for 1) potato variety, 2) cutting style, 3) skin presence, 4) grade level and 5) manufacturing line respectively. Usage of those importance factors minimizes all performance measures. The distance with respect to the order to be predicted can be measured for all historical orders. In particular, we can search for all orders that match the order to be predicted on all variables, i.e. the ones with distance equal to zero. The k-nearest neighbour interpolation is schematically displayed in Figure 6.



(a) Equal orders $\geq k$

(b) Equal orders $< k$

Figure 6: Visualization of the nearest neighbour based algorithm. The order to be predicted is located in the center of the circle. The numbers in the small circles represent group size. For some groups the factory recoveries $R_{i,j,k}$ are indicated.

The algorithm is based on comparison to the k nearest orders, with the possibility of varying k. The number of nearest neighbours in this research is set to $k = 10$. If the number of exactly matching orders is greater than or equal to k, all recoveries of those matching historical orders are used for prediction. The recoveries of all orders within the group are averaged. This is shown in Figure 6(a) where the number of equal orders is 26. However, if this number of matching orders is lower than k or even zero, the k closest neighbours are sought including the ones exactly matching. This is schematically displayed in Figure 6(b). In this case, instead of averaging the recoveries of those orders, a distance weighted average of the k recoveries is determined combining the nearest neighbour method with inverse distance weighting. If there are a few ($< k$) orders with distance zero, those orders are set to a distance of 0.01 to prevent dividing by zero in the following equation

$$\text{Predicted Recovery} = \frac{\sum_{i=1}^{k} R_i/D_i}{\sum_{i=1}^{k} 1/D_i}, \tag{12}$$

Using (12), in Figure 6(b), the recoveries of the four orders closer to the order to be predicted will contribute more than the orders in both groups of three. Notice that, if the importance factors of all predictor variables are equal, an order at distance 0.01 contributes a hundred times more than an order that differs on only one variable. The described process is translated into the following algorithm:

Algorithm 1 Calculating the predicted recovery for an order

```
procedure kNN-INTERPOLATION
    $NoO \leftarrow$ Number of historical orders
    $kNN \leftarrow$ Number of nearest neighbours taken into account for prediction
    $NoCA \leftarrow$ Number of categorical attributes
    $\gamma \leftarrow$ Array of weighting factors for categorical attributes with length $NoCA$
    for $i = 1 : NoO$ do
        for $j = 1 : NoCA$ do
            $Distance(i, j) = \gamma(j) * (Order(j) \neq Data(i, j))$
            $CatDistance(i) = CatDistance(i) + Distance(i, j)$
        if $CatDistance(i) = 0$ then
            $NoEO + 1$
            $CatDistance(i) = CatDistance(i) + 0.01$
    if $NoEO >= kNN$ then
        $kNN \leftarrow NoEO$
    $[MinDist, Idx] \leftarrow mink(CatDistance, kNN)$
    $Recoveries \leftarrow RecoveryData(Idx)$
    $PredictedRec \leftarrow \frac{\sum(1/MinDist*Recoveries)}{\sum(1/MinDist)}$
```

The performance measures of the best model obtained, after tuning the importance factors by means of cross-validation, are displayed in Table 4.

Table 4: Performance measures of the IDW-model.

Number of records	**Mean Absolute Error**	**Root Mean Square Error**	**Within variation**	**Within variation w/o singles**
3006	2.5628%	3.3640%	94.48%	99.93%

The algorithm can be extended to include numerical values. To do so, the distance measure D should be split in a numerical and categorical distance. If the traceability of the batch data is improved, it can be investigated whether numerical data leads to even more precise predictions.

To investigate whether the results above are obtained by means of the proper working of the model instead of over-fitting on the historical 3,006 orders, the model is cross-validated with different compositions of training and test groups. The training group in this case will constitute the basis of the model, in other words the orders used for building the predictions. The test group will not be used to build the model, only to validate the functioning. The performance measures applied on the test group for different compositions are displayed in Table 5. The compositions represent the percentages of orders used in both groups. For every composition, a random test and training group is generated ten times. Then, the performance measures are calculated ten times and the outcomes are averaged. As can be seen, even with a large test group that is not used to generate the model, the accuracy is reasonable. The accuracy increases when the number of orders that serve as the basis of the model becomes larger. Since the working of the model is now ratified, all 3,006 orders can be used as the basis of the model for prediction of the possible 82,350 combinations. These predictions can then be used to reallocate the raw material in Eqs. (1)−(7).

Table 5: Performance measures for different training and test groups calculated for the test group.

Training: Test ratio	**Training gr. size**	**Test gr. size**	**MAE**	**RMSE**	**Within 1.5σ**	**Within 1.5σ w/o singles**
60 : 40	1804	1202	3.3779%	4.5175%	93.01%	97.56%
70 : 30	2104	902	3.2716%	4.3119%	93.57%	98.60%
80 : 20	2405	601	3.3074%	4.4220%	94.01%	99.16%
90 : 10	2705	301	3.1618%	4.3363%	95.28%	99.31%
95 : 5	2856	150	3.0043%	3.9840%	95.33%	99.28%

5.2. Random Forest for Regression

Another suitable technique for prediction is the use of regression trees. A regression tree is a piecewise constant or piecewise linear estimate of a regression function, constructed by recursively partitioning the data and sample space [7]. It splits the data points based on the predictor values with the goal of minimizing the total error until a certain stopping criterion is reached. The predictor values are the same categorical predictors as used in the previous section. The ends of the tree, the leaves, contain the prediction values based on the weighted average of the data points in the leaf. This tree can then be used to predict the output of "new" combinations of input variables. Those combinations can be passed through the tree from top to bottom till they end up in a specific leaf.

However, a single regression tree has the tendency to be inaccurate because it predicts too general or because it is over-fitting on the available data. To decrease the mean squared error of the prediction, multiple trees can be generated and combined into a so-called random forest. The creation of a random forest is described in Algorithm 2.

Algorithm 2 The creation of a random forest

```
NoT ← Number of trees
C ← Number of categorical predictors
for $i = 1 : NoT$ do
    Create a bootstrapped dataset using random sampling with replacement
    while splitting criterion not violated do
        Determine residuals $r_i = y_i - \bar{y}$ of the node to be split, with $\bar{y}$
        the average of the node
        Assign the observations to a nominal bin variable $z_t$ according to the sign
        of the corresponding residuals
        Select $F = \text{int}(\log_2(C+1))$ random predictors
        Conduct curvature tests between each predictor in $F$ and the
        nominal bin variable $z_t$
        if All $\boldsymbol{p}$-values $\geq 0.05$ then do not split node $t$
        if Minimal $\boldsymbol{p}$-value $< 0.05$ then split node $t$ based on
        the corresponding predictor
        Set optimal cut point that yields the largest MSE reduction
```

The number of trees needs to be big enough for convergence of the mean squared error. For each tree in the regression forest a bootstrapped dataset is used to train the model. This implies that a set with the same size as the original set is generated by randomly picking samples from the original data set; it is allowed to pick a sample from the original data set multiple times. This results in unique regression trees, since the training data differs for every training process. The stopping criteria prevent splitting a node if either the node size is too small or the leaf size after splitting will be too small. Before every potential split a set of two predictors is randomly taken. The curvature test, a type of χ^2-test, is applied to determine the best predictor in the set to split a node. In that case, the best split predictor variable is the one that minimizes the significant $\boldsymbol{p}$-values (those less than 0.05) of curvature tests between each random predictor and the records in the node. The categorical variable that is used for splitting the node is called the splitting predictor. This is based on the variable selection method proposed by Loh [7]. Such a selection is robust to the number of levels in individual predictors. When the best splitting predictor is found, the node is split at the cut point that provides the largest reduction in mean squared error. To predict the output of "new" combinations of input variables, like using a single regression tree, the combinations are passed through all the trees in the forest. Finally, all the outcomes are averaged; this value represents the predicted factory recovery for the input combination. The process as a whole is based on the work of Breiman [6] and an example of the execution of the process is elaborated in Appendix B. The emergence of predictions using a random forest for regression is displayed in Figure 7.

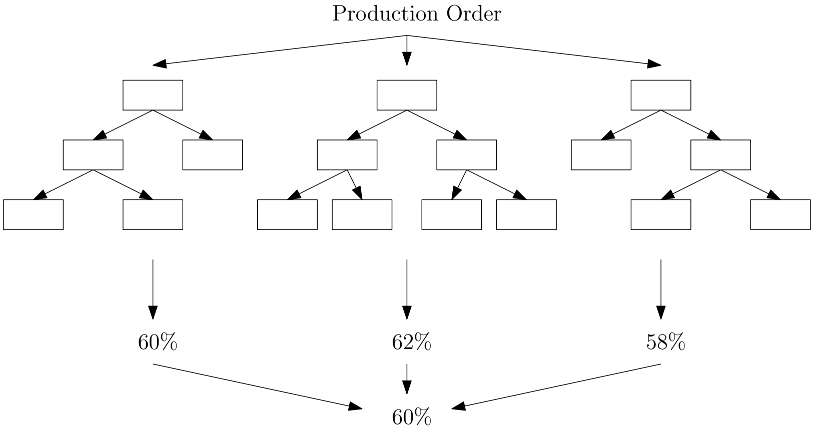


Figure 7: Visualization of prediction using a random forest for regression.

The technique described is applied on the 3,006 orders with the number of trees in the forest set to 500 to make sure the error converges to a minimal value. The created forest is then tested to predict the recoveries of the 3,006 orders. For this prediction, only the trees in the forest that were not based on the order to be predicted are used and their predictions are averaged. For the creation of a single tree not all orders are used due to the use of a bootstrapped data set. Those orders are called the Out-of-Bag orders and using them for prediction is a validation method to demonstrate the working of the model and to show that the results are not based on an over-fitted model. The calculated performance measures are displayed in Table 6. For prediction of the possible 82,350 combinations all trees in the generated forest can be used since the working has been validated.

Table 6: Characteristics random forest with categorical predictors only.

Number of records	**Mean Absolute Error**	**Root Mean Square Error**	**Within variation**	**Within variation w/o singles**
3006	3.1596%	4.1838%	91.72%	97.01%

Besides using the predictions of both techniques to solve the LP they can also be directly used by an end user. When the predictions are captured in an organized tool, the end user can determine how much his/her proposed production order is going to yield by selecting a combination of product, potato type and manufacturing line.

6. Results

Now, since it is possible to predict factory recovery using both techniques, the deployment phase can take place. The predicted recoveries can be used to reallocate the potatoes to product for the 3,006 orders to minimize potato input. When the linear program from Eqs. (1)−(7) is solved for those 3,006 orders, it is possible to find out the possible potato savings that could have been made. As stated before, 50 potato types, 183 product types and 9 manufacturing lines occur in those orders; $N = 50$, $M = 183$ and $L = 9$. The amount of available potatoes per type is determined by adding up the potatoes used for the historical orders per variety. In the same way the amounts of product to be generated are determined. The LP is solved using Gurobi Optimization software in Matlab in 0.9 seconds. The decision variables are the fractions which minimize potato use, yet satisfy all constraints. These fractions can be seen as the part of the total amount of a potato type available used for the production of a specific product on a specific line. To do so, all variable values are translated into numerics. So, $x_{9,63,2}$ is the fraction of potato type 9 used for the production of product 63 on manufacturing line 2, which corresponds to producing 9x9mm SPX Skin-Off Fries out of Bintje Field Ware on manufacturing line 2. Using the recoveries predicted with nearest neighbour interpolation the determined optimal fractions lead to a minimized potato use of 1.76 million tonnes. The amount of potatoes initially used to produce the orders is equal to 1.85 million tonnes, so a saving of 90 thousand tonnes (4.9%) is obtained. Using the recoveries predicted with the random forest technique a minimized potato use of 1.80 million tonnes is found; a saving of 50 thousand tonnes (2.7%). Figure 8 shows the use of material per variety expressed in the percentage of the total historical supply. This is calculated using

$$x_i = \frac{\sum_{j=1}^{M} \sum_{k=1}^{L} S_i x_{i,j,k}}{\sum_{i=1}^{N} S_i}, \qquad i = 1, 2, ..., N. \tag{13}$$

The reallocated schemes use mainly the same potato varieties as the allocation scheme executed by the manufacturer. However, the allocation scheme using kNN-interpolation mainly saves potatoes of type Innovator Field Ware and Premiere Field Ware with respect to the historical allocation. The optimal scheme is based on 228 active fractions, from which the factory recoveries of 150 combinations were not known before. The RF-based scheme on the other hand mainly saves on the use of Ramos Field Ware. In this case the scheme is based on 223 active fractions, from which 118 combinations were not executed before.

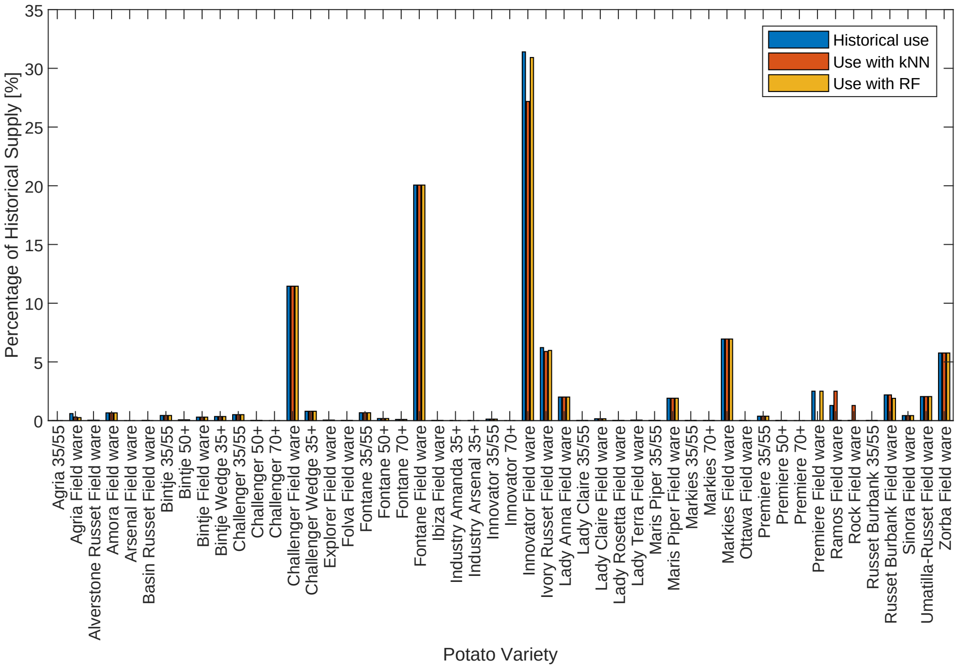


Figure 8: Use of the historical potato stock per variety for a) historical allocation, b) allocation using nearest neighbour interpolation and c) allocation using random forests.

After that, the program is used to determine what would have been an optimal potato purchase strategy. Therefore, it is assumed that there is an infinite stock of raw material varieties to determine which varieties are used to produce the 3,006 orders if they were available. Solving the program with the kNN-interpolation based prediction and the unlimited supply led to a potato use of 1.73 million tonnes. This is an additional saving of 24 thousand tonnes with respect to the determined kNN-interpolation based optimal policy with historical supply and with respect to the allocation executed by the manufacturer a saving of 115 thousand tonnes is obtained. Solving the program with the RF-based predictions and again unlimited supply led to a potato use of 1.78 million tonnes. This is an additional saving of 24 thousand tonnes with respect to the previous determined RF-based optimal policy and a saving of 74 thousand tonnes with respect to the historical scheme. The varieties used to obtain these results are displayed in Figure 9. Again, those are expressed in percentages of the total 1.85 million tonnes used by the manufacturer to produce the orders using (13). Remarkable is the drop in use of Innovator Field Ware when unlimited alternatives are available. The nearest neighbour interpolation based scheme prefers the use of Markies Field Ware to replace Innovator Field Ware, but the random forest based scheme prefers using Fontane 70+ and Lady Anna Field Ware. Differences in preferred varieties are due to differences in predicted factory recoveries which arise from natural variation in the executed orders.

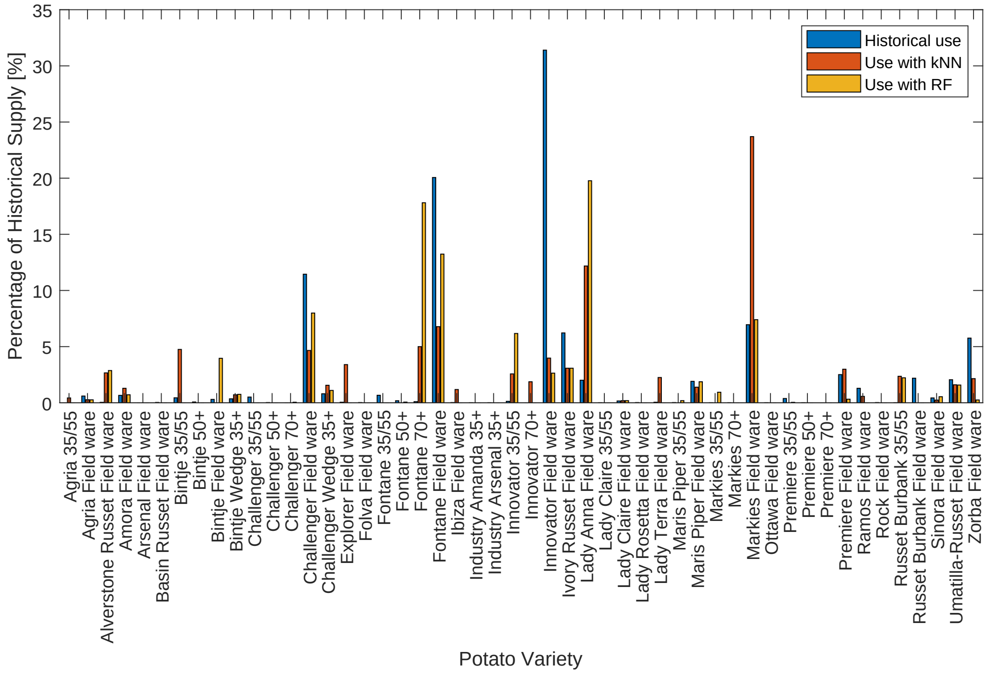


Figure 9: Percentage used of the historical potato stock with unlimited supply per variety for a) historical allocation, b) allocation using nearest neighbour interpolation and c) allocation using random forests.

6.1. Sensitivity Analysis

To demonstrate the importance of correct information when solving the LP several scenarios tests are performed. By means of the one-factor-at-a-time method several scenarios are plotted against the historical situation. Every time in the initial reallocated optimal program one factor is varied. The improvements with respect to the historical situation are displayed in Figure 10. The first set of bars displays the improvement made using Eqs. (1)−(7) and the data provided by the manufacturer. The other sets of bars display the adapted scenarios. Naturally, the program with unlimited supply provides the biggest improvement. Production with an overall capacity decrease of 30% is still possible. This is interesting to see if for some reason the provided capacity data is lower than expected. Also with a small decrease in factory recovery, allocation is still feasible. Even if the factory recovery of all orders is 0.5% lower than predicted, an improvement with respect to the historical allocation is obtained. However, if factory recovery is decreased further, the production amount constraint is violated for particular product types. Thus, it is impossible to generate the desired product amounts.

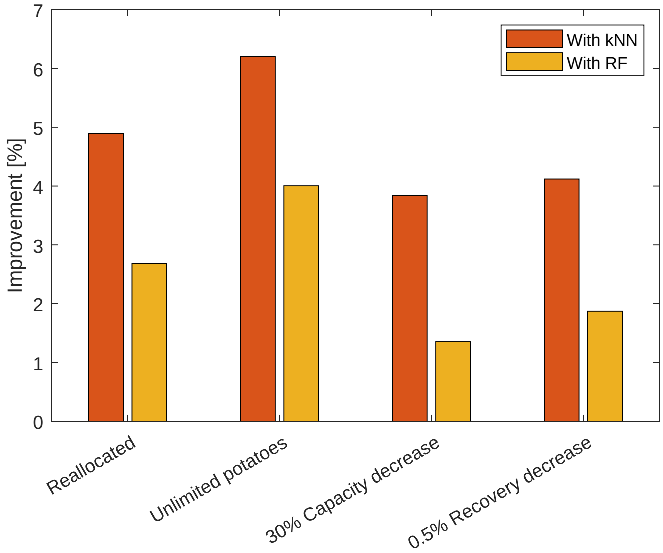


Figure 10: Percentage improvement with respect to the historical allocation. The tested scenarios are 1) the reallocation using Eqs. (1)−(7), 2) with unlimited potato supply, 3) with an overall decrease in capacity and 4) with an overall decrease in recovery.

Differences exist between the results of both techniques. The potential savings with respect to the historical allocation of the kNN-based schemes are always higher than the savings of the RF-based schemes. Besides that, in the scenario with unlimited potato supply for both schemes different potato types are preferred. A possible explanation of these differences may be the natural variation of the orders on which the predictions techniques are applied. This natural variation affects the predictions which is also visible in their averages. The nearest neighbour interpolation based recoveries are 61.9% on average, whereas the random forest based predictions have an average of 60.6%.

The scenario tests can be supplemented with sensitivity analysis of the optimal LP solutions. Commercial linear programming systems usually provide sensitivity analysis results and so does Gurobi. Gurobi provides so-called shadow prices which indicate the change in the optimal solution if the right hand side of the constraints is slightly adapted. For example, one additional kilogram of demand D of product type 18 leads to 1.96 kilogram additional potato use. The magnitude of the shadow prices indicates which components in the LP have a major influence on the optimal solution, which are therefore desired to be of high accuracy, and which components are less influential. Important constraints are defined as the ones which contribute most to the optimal potato use. In this sensitivity analysis the influences of the potato availability constraints, the product demand constraints and the line capacity constraints, respectively (2), (3) and (4) in Eqs. (1)$-$(7), on the optimal solution are investigated. The shadow prices of these types of constraints all represent something different. The shadow prices of the potato availability state how much the optimal potato use could be decreased if there was a bigger stock of a type available. The shadow prices of the product demand constraints indicate how much additional potatoes are needed to produce an additional kilogram of product. The shadow prices of the line capacity show how much the optimal potato use is decreased if an additional hour on a manufacturing line becomes available. For the three types of constraints the five constraints that have the most impact on the optimal solution when changed are displayed in Table 7.

Table 7: Most important constraints with regard to the optimal solution following from sensitivity analysis sorted by importance.

	Potato Availability		**Product Demand**		**Line Capacity**	
#	**kNN**	**RF**	**kNN**	**RF**	**kNN**	**RF**
1.	46	46	112	110	3	2
2.	29	49	18	112	2	3
3.	4	29	75	75	-	-
4.	33	48	111	111	-	-
5.	30	4	80	108	-	-

The results in Table 7 indicate that for example the accuracy of the potato availability of type 46, Russet Burbank, is rather important. Besides that it states that increasing the stock of this type leads to the biggest decrease in optimal potato use. Also the product demand of type 112, 6x6mm RPX Skin-off fries, needs to be determined correctly. An increase of demand of this type leads to the largest potato use increase, which is not desired. Notice that for the line capacity constraints only two lines are influential. This is due to the fact that the other constraints are not active yet; there is still capacity left for production. Indirectly the things stated above lead to the importance of predicting recoveries for those varieties, products and lines correctly.

The shadow prices found using Gurobi can also be used to express sensitivity in the left hand side of the LP, the constraint matrix. For infinitesimal small adaptations to the constraints, an approximation of the shadow prices of the left hand side components is found by multiplying the shadow prices of the right hand side with the optimal solution [12]. The LP solution is rather sensitive with respect to the predictions for factory recovery. This is shown in the fourth set of bars in Figure 10, but also in the difference between the results of the kNN-based and RF-based schemes. By means of the sensitivity analysis for the left hand side, it is found which recoveries are important to be predicted correctly. Those recoveries are displayed in Table 8.

Table 8: Most influential factory recoveries with regard to the optimal solution following from sensitivity analysis sorted by importance.

	Factory	**Recovery**		**Factory**	**Recovery**
#	**kNN**	**RF**	#	**kNN**	**RF**
1.	$R_{27,159,5}$	$R_{27,159,4}$	6.	$R_{28,115,9}$	$R_{14,1,1}$
2.	$R_{21,36,4}$	$R_{27,113,1}$	7.	$R_{28,109,9}$	$R_{28,115,9}$
3.	$R_{27,113,3}$	$R_{21,51,6}$	8.	$R_{14,155,1}$	$R_{28,109,9}$
4.	$R_{27,150,3}$	$R_{27,151,4}$	9.	$R_{27,114,8}$	$R_{38,155,4}$
5.	$R_{38,1,1}$	$R_{27,150,3}$	10.	$R_{38,158,4}$	$R_{27,114,8}$

The recoveries $R_{i,j,k}$ represent the recovery belonging to the production of product j out of potato type i on manufacturing line k. Notice that a lot of those recoveries are regarding potato type 27, Innovator Field Ware. Since this potato type is the one with the biggest stock, according to historical data, accurate prediction of recovery for this type is very important. Besides that, there are several matching combinations, even though different techniques are used, that are important for the optimal solution. The presence of similar important constraints using both techniques supports the presumption of the correctness.

7. Extensions

The program in Eqs. (1)−(7) can be extended to incorporate costs. Since the objective is to reduce mass waste, the optimal solution can use more expensive potatoes over cheaper ones, as long as the mass used is minimal. To overcome this, the objective needs to be adapted to minimize costs instead of mass use. This leads to the introduction of purchase expenses E_i, dependent on the type of potatoes expressed in price per kilo. The following program is drafted:

$$\begin{aligned}
\text{minimize} \quad & \sum_{i=1}^{N}\sum_{j=1}^{M}\sum_{k=1}^{L} E_i S_i x_{i,j,k} & & \\
\text{subject to} \quad & \sum_{j=1}^{M}\sum_{k=1}^{L} x_{i,j,k} \leq 1, & i = 1, 2, ..., N & \\
& \sum_{i=1}^{N}\sum_{k=1}^{L} S_i R_{i,j,k} x_{i,j,k} \geq D_j, & j = 1, 2, ..., M & \\
& \sum_{i=1}^{N}\sum_{j=1}^{M} S_i R_{i,j,k} x_{i,j,k} \leq C_k, & k = 1, 2, ..., L & \qquad (14) \\
& \sum_{i \in \mathbf{I}_j}^{N}\sum_{k=1}^{L} x_{i,j,k} = 0, & j = 1, 2, ..., M & \\
& \sum_{k \in \mathbf{K}_j}^{L}\sum_{i=1}^{N} x_{i,j,k} = 0, & j = 1, 2, ..., M & \\
& x_{i,j,k} \geq 0, & \text{for all } i, j, k &
\end{aligned}$$

Now the extended LP in (14) can be solved, minimizing costs of potatoes instead of minimizing the amount of potatoes used. Therefore, the expenses on potato purchase should be added. Those are added as the purchase costs per kilo of a potato type E_i. Solving this system, again using Gurobi Optimization software in Matlab, took 0.3 seconds. The purchase of potatoes used to produce the 3,006 orders costed €220M. Solving the extended program using the nearest neighbour interpolation based predictions leads to purchasing costs of €209M, which is a saving of €11M (5.14%). Solving the program using the random forest based predictions leads to purchasing costs of €214M, which is a saving of €6M (2.81%).

To complete the allocation process, also transportation costs should be considered. Since production on a particular line can be beneficial, it may be rewarding to transport the available potatoes from one place to another. However, if the benefits with respect to recovery are minimal and transport is an expensive operation, production without transport should be considered. Besides that, the storage location of the finished product could also lead to additional transportation costs. If customers are close to the storage location, transportation costs will be considerably lower. Thus, if the product is generated on a line far from the storage location additional transportation costs occur. To find out whether production benefits outweigh transportation costs, the costs are included in the form of TCT, transportation costs from potato storage **to** the manufacturing line and TCF, transportation costs **from** the manufacturing line to the product storage location. This adapted system is given in (15).

$$\begin{aligned}
\text{min} \quad & \sum_{i=1}^{N}\sum_{j=1}^{M}\sum_{k=1}^{L}\sum_{a=1}^{P}\sum_{b=1}^{Q}(E_iS_i + TCT_{a,k} + TCF_{k,b})x_{i,j,k}^{a,b} & & \\
\text{s.t.} \quad & \sum_{j=1}^{M}\sum_{k=1}^{L}\sum_{a=1}^{P}\sum_{b=1}^{Q} x_{i,j,k}^{a,b} \leq 1, & i &= 1,2,...,N \\
& \sum_{i=1}^{N}\sum_{k=1}^{L}\sum_{a=1}^{P}\sum_{b=1}^{Q} S_iR_{i,j,k}x_{i,j,k}^{a,b} \geq D_j, & j &= 1,2,...,M \\
& \sum_{i=1}^{N}\sum_{j=1}^{M}\sum_{a=1}^{P}\sum_{b=1}^{Q} S_iR_{i,j,k}x_{i,j,k}^{a,b} \leq C_k, & k &= 1,2,...,L \qquad (15) \\
& \sum_{i\in\mathbf{I}_j}^{N}\sum_{k=1}^{L}\sum_{a=1}^{P}\sum_{b=1}^{Q} x_{i,j,k}^{a,b} = 0, & j &= 1,2,...,M \\
& \sum_{k\in\mathbf{K}_j}^{L}\sum_{i=1}^{N}\sum_{a=1}^{P}\sum_{b=1}^{Q} x_{i,j,k}^{a,b} = 0, & j &= 1,2,...,M \\
& x_{i,j,k}^{a,b} \geq 0, & &\text{for all } i,j,k,a,b
\end{aligned}$$

The fraction $x_{i,j,k}$ as in Eqs. (1)−(7) and (14) is extended too to $x_{i,j,k}^{a,b}$ to indicate where the potatoes were stored (a) and where the product has to be stored (b). Since data on potato and product storage location is missing in the current data, this program can not be solved immediately. It leaves an interesting problem, since not only transport within the current network can be considered, but there also could be thought of adapting the network. For example, one can think about the number of storage locations and where to position them.

8. Conclusions

A raw material saving approach for a potato product manufacturer is provided which requires few radical measures with respect to the manufacturing process. By means of collected data it is possible to improve the allocation of potatoes to end product and manufacturing lines to minimize raw material use. The raw material used for 3,006 considered production orders from July 2016 till July 2018 is reallocated in a way that equal amount of product is produced, but less potatoes are used. The allocation is based on factory recovery and user-specified constraints. The application of nearest neighbour interpolation and random forests for regression on historical production orders led to accurate prediction of factory recoveries of all combinations of material, product and line that can occur at the manufacturer. It is showed that on average the approach would have saved 3.8% with respect to potato mass input and 4.0% with respect to purchasing costs over the stated period. Besides that, in combination with sales forecasts of end product recommendations can be made on how to distribute the available material or even ideal purchasing policies can be opted for the time to come.

From sensitivity analysis, it follows that allocation is highly dependent on the accuracy of the factory recovery prediction. Using the nearest neighbour interpolation model the predictions had a mean absolute error of 2.6%. Those predictions were based on 1) potato variety, 2) cutting style of the product, 3) skin presence, 4) grade level of the product and 5) manufacturing line to produce on; categorical variables that are known before production starts. Using a random forest model and the same five categorical predictors a mean absolute error of 3.2% is obtained.

Allocation is a process that occurs within other manufacturing companies too, so the proposed approach could be used here as well. To do so, the availability of raw material (types), the desired amount of product or expected sales and the capacity of the lines should be known. All those variables should be determined for the same time period, otherwise allocation would not be adequate. The constraints on the program can be extended if necessary for the considered process, for example if a product needs to be made out of multiple materials which can have different varieties (e.g. beer).

9. Further Research

When a real sustainable impact is desired, with or without considering financial costs, the proposed framework should be used to minimize energy footprint or a similar measure. The purchase network can be optimized leading to better targeted contracting to obtain higher yield from the raw material and to prevent unnecessary production and transport emissions. A measure should be introduced to determine the ecological impact from the moment the crops are planted till the product is delivered to the customer. Therefore, traceability is very important, which is often a bottleneck in data mining projects like these.

If traceability is improved, numerical variables like potato length and underwater weight can be used for prediction. These predictors could lead to improved prediction accuracy, leading to more reliable allocation results. As showed by means of the sensitivity analysis, factory recovery has a noticeable influence on the optimal allocation schemes.

Besides that, if the processing part is not longer considered as a black box, better targeted recommendations can be made with respect to the manufacturing lines of the company. In the current setting it is unclear to see where losses occur and if they could have been prevented. Besides that, it is also not possible to track down the efficient operations.

Another step can be made in scheduling the production orders. The allocation framework only considers a fixed time period and assumes that potatoes are available and remain available at all times during that period. However, degradation of the material occurs leading to deterioration in potato quality and some potato varieties become available only later in the harvesting process. Therefore, a time-dependent scheduling approach should be introduced taking the optimized allocation into account for optimal yield.

References

[1] Gustavsson, J., Cederberg, C., Sonesson, U., van Otterdijk, R., Meybeck, A. (2011). *Global food losses and food waste - Extent, causes and prevention.* Rome

[2] Henningsson, S., Smith, A., & Hyde, K. (2001). Minimizing material flows and utility use to increase profitability in the food and drink industry. *Trends in Food Science & Technology, 12*(2), 75-82. https://doi.org/10.1016/S0924-2244(01)00052-8

[3] Papargyropoulou, E., Lozano, R., Steinberger, J.K., & Wright, N. (2014). The food waste hierarchy as a framework for the management of food surplus and food waste. *Journal of Cleaner Production, 76*, 106-115. https://doi.org/10.1016/j.jclepro.2014.04.020

[4] Zhong, R.Y., Xu, X., Klotz, E., & Newman, S.T. (2017). Intelligent Manufacturing in the Context of Industry 4.0: A Review. *Engineering, 3*(5), 616-630. https://doi.org/10.1016/J.ENG.2017.05.015

[5] Chapman, P., Clinton, J., Kerber, R., Khabaza, T., Reinartz, T., Shearer, C. & Wirth, R. (2000). *CRISP-DM 1.0 Step-by-step data mining guide*. The CRISP-DM consortium.

[6] Breiman, L. E. O. (2001). Random Forests. *Machine Learning, 45*(1), 5-32.

[7] Loh, W. (2002). Regression Trees with Unbiased Variable Selection and Interaction Detection. *Statistica Sinica, 12*, 361-386.

[8] Kotsiantis, S. B. (2007). Supervised Machine Learning: A Review of Classification Techniques. *Informatica, 31*(3), 249-268.

[9] Jerez, M., Molina, I., Garca, P.J., Alba, E., Ribelles, N., Franco, L., & Martn, M. (2010). Artificial Intelligence in Medicine Missing data imputation using statistical and machine learning methods in a real breast cancer problem. *Artificial Intelligence in Medicine, 50*(2), 105-115. https://doi.org/10.1016/j.artmed.2010.05.002

[10] De'ath, G., & Fabricius, K.E. (2000). Classification and regression trees: A powerful yet simple technique for ecological data analysis. *Ecology, 81*(11), 3178-3192. https://doi.org/10.1890/0012-9658(2000)081[3178:CARTAP]2.0.CO;2

[11] Joseph, V.R., & Kang, L. (2012). Regression-Based Inverse Distance Weighting With Applications to Computer Experiments. *Technometrics, 53*(3), 254-265. https://doi.org/10.1198/TECH.2011.09154

[12] Freund, R.M. (1985). The sensitivity of a linear program solution to changes in matrix coefficients. *Sloan Working Paper No. 1532-84*, Massachusetts Institute of Technology, Cambridge, Massachusetts.

A. Appendix: Data set sample

Table 9: A snippet of the complete data set with the most important variables. The variables “Plant Name” and “Manufacturing Line” are combined into the new variable “Line”. The columns “Potatoes input” and “Product Output” are redundant, because those are captured together by “Factory Recovery”.

Order	Material	Factory Recovery	Variety Cluster	Variety	Short Code	Cutting Style	Skin	Line	Grade
1190708	44954	0.660792	Variety Cluster 300FA	Premiere Field ware	CKF10	9x9 mm	No	0040-1	RFA
1190710	45033	0.622392	Variety Cluster 230LF	Amora Field ware	WW3	11x11 mm	No	0040-1	RLF
1190997	60723	0.640357	Variety Cluster 264LF	Premiere Field ware	LWS67	19x19x23	No	0020-3	RFW
1190998	34249	0.758158	Variety Cluster 252LF	Premiere Field ware	LWS12N	7/16” CrissCuts	No	0020-3	SFW
1191000	52753	0.75095	Variety Cluster 252LF	Premiere Field ware	D24	7/16” CrissCuts	No	0020-3	SFW
1191027	52792	0.593275	Variety Cluster 122PXLF	Zorba Field ware	F64	9x9 mm	No	0020-1	RPX
1191029	46626	0.696097	Variety Cluster 251LF	Premiere Field ware	SC06	1/4” Twister	No	0020-3	SFW
...									

B. Appendix: Random Forests for Regression

To give a better impression of the generation of a random forest for regression an example is elaborated. A small data set is shown in Table 10. Usually a much bigger data set is used for creation of regression forests. The number of predictors C in this case is five and the last column represents the target value.

Table 10: Example data set for the creation of a regression forest.

Potato variety	**Manu. line**	**Cutting size**	**Potato skin**	**Grading level**	**Factory recovery**
Innovator	0020-1	9x9mm	No	RPX	63%
Challenger	0020-1	6x6mm	No	RPX	52%
Challenger	0030-1	9x9mm	No	SFW	61%
Bintje	0020-2	6x21mm	Yes	SLF	72%

Now, to create the first regression tree in the forest, a bootstrapped dataset is taken. This implies that a set with the same size as the original set is generated picking samples from the original data set. It is allowed to pick a sample from the original data set multiple times. The bootstrapped dataset for the example is displayed in Table 11. Notice that the second sample from the original sample ended up twice in the bootstrapped dataset.

Table 11: Bootstrapped dataset for the creation of the first regression tree.

Potato variety	**Manu. line**	**Cutting size**	**Potato skin**	**Grading level**	**Factory recovery**
Innovator	0020-1	9x9mm	No	RPX	63%
Challenger	0020-1	6x6mm	No	RPX	52%
Challenger	0030-1	9x9mm	No	SFW	61%
Challenger	0020-1	6x6mm	No	RPX	52%

Now, a regression tree needs to be generated based on this bootstrapped dataset. However, instead of splitting based on all five predictors, at every split the first integer less than $\log_2(C+1)$ of predictor values is taken to split upon. In the example $C = 5$, so $\log_2(C+1) \approx 2.59$, so the first integer smaller than this is two. For the first split the two random variables to possibly split upon are “Potato Variety” and “Potato Skin”. Curvature tests needs to be conducted to determine the optimal split variable. Curvature tests are χ^2-like tests to test the hypothesis that there is no relation between the predictor and the records in the node. The predictor with the lowest $\boldsymbol{p}$-value for this test is used to split the node. A very small chi square test statistic ($\boldsymbol{p}$-value) means that there is a strong relationship between the predictor and the points in the node. Suppose the first split is made based on "Potato Variety" for this example, which means the orders in the node are divided based on potato variety. Then two predictor values are taken again to make the second split. This continues until the splitting criteria are violated. The regression values are the averages of the factory recovery of the samples in an end node called a leaf. The first tree is now created.

After that, a new bootstrapped dataset is taken and a second tree is generated; this process is typically repeated 100 times. To predict the factory recovery of a new combination of variables, this combination could be ran through the multiple regression trees. It ends up in 100 different end nodes, from which the average value can be taken as the final response value.

TU/e Technische Universiteit
Eindhoven
University of Technology

Declaration concerning the TU/e Code of Scientific Conduct for the Master's thesis

I have read the TU/e Code of Scientific Conduct[i].

I hereby declare that my Master's thesis has been carried out in accordance with the rules of the TU/e Code of Scientific Conduct

Date

04-06-2019

Name

Nick Willems

ID-number

0851016

Signature

Nick

Submit the signed declaration to the student administration of your department.

[i] See: http://www.tue.nl/en/university/about-the-university/integrity/scientific-integrity/
The Netherlands Code of Conduct for Academic Practice of the VSNU can be found here also.
More information about scientific integrity is published on the websites of TU/e and VSNU