Computers and Electronics in Agriculture 127 (2016) 467–474

Contents lists available at ScienceDirect

Computers and Electronics in Agriculture

journal homepage: www.elsevier.com/locate/compag

Original papers

Soybean varieties portfolio optimisation based on yield prediction

Oskar Marko *, Sanja Brdar, Marko Panic, Predrag Lugonja, Vladimir Crnojevic

BioSense Institute, Dr Zorana Djindjica 1, 21000 Novi Sad, Serbia

ARTICLE INFO

Article history:
Received 31 March 2016
Received in revised form 5 July 2016
Accepted 9 July 2016

Keywords:
Yield prediction
Seed selection
Weighted histograms
Portfolio optimization
Convex optimization

ABSTRACT

One of the biggest problems in agriculture is concerned with seed selection. Wrong choice of seed variety cannot be compensated with fertilisation, spraying or the use of mechanisation later in the season. The purpose of this work was to design the strategy for selecting soybean varieties that should be planted on the test farm in order to maximise yield in the following season, based on the knowledge acquired from heterogeneous historical data. We propose weighted histograms regression to predict the yield of different varieties and compare our method to conventional regression algorithms. Based on the predicted yield, we perform portfolio optimisation to come up with the optimal selection of seed varieties that is to be planted. Presented algorithms and results were produced within the Syngenta Crop Challenge.

1. Introduction

The world's growing food demand (Godfray et al., 2010; Ash et al., 2010) challenges seed industry to develop and improve seed varieties, but also challenges farmers to select appropriate seeds among hundreds of varieties available nowadays (Sperling et al., 2014; McGuire and Sperling, 2016). What farmers would certainly need is a portfolio of seed varieties, customised for the environmental conditions at their farm, which would maximise the yield and reduce the insecurity that comes from its variability (Hanson, 2013). Such a targeted solution is important for both traditional (Yengoh, 2012; Louette et al., 2000) and precision agriculture, where the decisions are made locally, on the smallest possible scale (Gassner et al., 2013). In general, there are numerous parameters that influence crop yield. Most prominent are climate and weather conditions, soil type, seed variety and land management, but in the end, it is their complex interaction that determines the yield.

In order to make the decision which seed varieties would be suitable for the given parcel and its environmental parameters, it is necessary to predict their yields. There is an increasing number of scientific researches dealing with yield prediction of various types of crops, fruit and vegetables. Some are based on image processing like in (Pantazi et al., 2016; Liakos et al., 2015), where yield was predicted using NDVI extracted from satellite images and images acquired by a handheld camera. Another approach is to analyse the physical properties of plants, such as height, grain weight and peduncle length (Romero et al., 2013), number of flowers on apple trees (Aggelopoulou et al., 2011) or chlorophyll content measured with SPAD (Saruta et al., 2013). Weather data can also serve as input for yield prediction (Marinković et al., 2009; Brdar et al., 2011; Gonzalez-Sanchez et al., 2014). For example, rainfall in May and a lot of sunshine in June can positively affect the yield of wheat in Serbia, whereas dry spring and extremely hot June can affect it negatively.

The problem with seed selection is that, no matter how successful they may be, none of the aforementioned in-season methods can be applied. It is impossible to know crop vigor, plant height or even weather conditions for the next year. However, the condition of soil does not change dramatically one year after another. It has been shown that content of organic matter, phosphorus, calcium and other compounds in the soil, as well as its pH value, are good indicators of the amount of yield (Drummond et al., 2003). Furthermore, yield can be also estimated based on the ratio of clay, silt and sand, and soil's shallow electrical conductivity (Papageorgiou et al., 2013).

As for the algorithms used for yield prediction, most common ones are artificial neural networks (ANNs) (Pantazi et al., 2016; Drummond et al., 2003; Freitas et al., 2009; Uno et al., 2005; Kaul et al., 2005), multiple regression (Drummond et al., 2003; Kaul et al., 2005) and regression trees (Romero et al., 2013; Marinković et al., 2009). In this work we propose a novel approach to yield prediction – weighted histograms regression (WHR). We approximate the yield probability density function (PDF) at the test farm by forming a histogram of yield, whose entries are weighted according to similarity between test and training farms.

Weighted histograms are not completely new. They have already been used in image processing for motion tracking

* Corresponding author.
E-mail address: oskar.marko@uns.ac.rs (O. Marko).

http://dx.doi.org/10.1016/j.compag.2016.07.009
0168-1699/

(Comaniciu et al., 2003), where an object's feature PDF needs to be calculated. Pixels in the centre of an object are more reliable and thus are attributed with a higher weight. Peripheral pixels are less reliable due to occlusion and interference from the background, and are thus taken with a lower weight. Also, in object recognition, target objects are compared to objects from the database by colour histograms. Since colour is susceptible to changes caused by varying illumination, similar colours are also taken into account – the more similar they are, the more they will contribute to the histogram (Jia et al., 2006).

Yield prediction is just a step towards seed selection. Having known the values of yield predicted for each seed variety, portfolio optimisation theory comes into play. It is a well established theory, originally used for choosing the right portfolio of investments on stock market, which would maximise the return and minimise the risk (Markowitz, 1952). Lately, there have been some examples of its usage in agriculture, as well. It is usually employed in seed variety selection (Nalley et al., 2009; Nalley and Barkley, 2010; Barkley et al., 2010), where predicted yield corresponds to financial return (Freitas et al., 2009), but there are also cases of its use for e.g. irrigation decision-making in condition of reduced water availability (Paydar and Qureshi, 2012), forest planning under the effects of climate change (Dragicevic et al., 2016) and for selecting optimal mix of tree families (Weng et al., 2013). It is always a good strategy to grow plants that respond differently to different environmental conditions and thus statistically better cope with weather unpredictability (Di Falco, 2012). This is especially important for ensuring yield stability in low-income nations and increasing drought and pest tolerance of crops (Barkley et al., 2010).

Whereas high prediction accuracy has been achieved only with classification of the yield into categories, such as low, medium and high (Romero et al., 2013; Saruta et al., 2013; Papageorgiou et al., 2013) and with in-season predictions (Marinković et al., 2009; Brdar et al., 2011; Kaul et al., 2005), we show that it is possible to achieve a high accuracy prediction for one year in advance by using the weighted histograms regression approach. By using this method along with convex optimisation and portfolio optimisation theory it is possible to select a portfolio of seeds, which maximises the yield.

2. Data

Algorithms and related results presented in this paper have been in part developed within the Syngenta Crop Challenge (Syngenta Crop Challenge, 2016), where competitors were provided with necessary historical data about soil, yield and soybean varieties used. The dataset contained 34,212 entries with any of 180 seed varieties planted on one of 120 farms located in the American Midwest (Fig. 1). The varieties were represented with anonymised IDs – v_i, where i took 180 values within the range from 1 to 210.

Season, geographic location, soil properties, common practice and other related parameters were given as features and are listed in Table 1. The sources of data were Syngenta's internal database, ISRIC (World Soil Information) (Hengl et al., 2014), CONUS (Soil Information for Environmental Modeling and Ecosystem Management) (Miller and White, 1998), NASS (United States National Agricultural Statistics Service) (Boryan et al., 2011) and FAO (United Nation's Food and Agriculture Organisation) (FAO, 2016). Some features were contained in datasets of two independent sources. Values from multiple sources were treated as separate features and were all used for prediction.

In the preprocessing phase, we detected that there were multiple entries with the same seed variety planted on the same farm in the same year, but with a different value of yield. We merged them and used only the average yield value accordingly, leaving 32,120 entries. In order to avoid bias and provide reliable yield prediction results for one season, we further split the set into training (seasons 2008–2013) and test dataset (season 2014), with 21,121 and 10,999 entries, respectively. We used only year 2014 as the test dataset to maintain the time causality of our approach. Each year the overall yield gets bigger because of the better mechanisation, pesticides and fertilisers used, as well as other improvements in agricultural production and it was crucial to capture this trendline. In this manner we tried predicting yield in previous years as well. However, the available training dataset reduced dramatically for each preceding year. The number of training samples was insufficient to successfully predict the yield in years before 2014.

3. Methodology and theory

3.1. Prediction using weighted histograms

In order to get the idea about a complexity of the given problem, we used the most straightforward approach by checking the correlation between the yield and individual features, but we did not get any meaningful results. There was no direct link between any of the parameters and the yield. Consequently, we proposed a novel method with the underlying principle that the agricultural system is determinative, i.e. with the same environmental conditions, soil characteristics and seed varieties, different farms give the same yield. In other words, when features of any two farms are compared, the more similar the features are, the more likely it is for the farms to have similar yield. The detailed description of the method follows.

The goal of Syngenta Crop Challenge was to choose up to five soybean varieties that should be planted on the so-called "Evaluation Farm" to maximise the yield. Firstly, we chose a soy variety whose yield we wanted to predict at a test farm. The process was repeated for all available varieties. In the following example, the evaluation farm was denoted as F_E and the variety of interest as v_x. Let us assume that there were five instances of planting v_x in the training dataset. Although this particular variety can be planted on the same farm throughout different years, we can assume without the loss of generality that it was planted on five different farms (F_1 to F_5) (Fig. 2).

Next, we considered the similarity between environmental conditions and other properties at training farms where v_x was planted and related properties at the evaluation farm. They were compared according to individual features – one feature at a time. Let us denote an arbitrary feature according to which the similarity was measured as f_i. The example in Fig. 3 shows values of the given feature at different farms, where the superscript indicates the farm it is related to.

Accordingly, distances of training farms from evaluation farm in the feature's space are shown in descending order in Table 2.

The yield at evaluation farm was more likely to resemble the yield at farms whose value of f_i was closer to f_i^E. Likewise, we could not expect the yield at the evaluation farm to correspond to the yield of a training farm if they had completely different f_is. Another way of explaining this is to view the training farms as advisers, who give their opinion about the yield at the evaluation farm. However, their opinions were not equally important. Opinions of training farms whose f_i was closer to f_i^E were taken with a higher significance than the opinions of those farms whose f_i was far away from f_i^E. Furthermore, a farm's opinion was simply the value of its own yield. It was as if the training farms were telling the evaluation farm that it would have the same yield as them, but the evaluation farm valued their opinions according to how far they were with respect to the given feature. In order to quantify

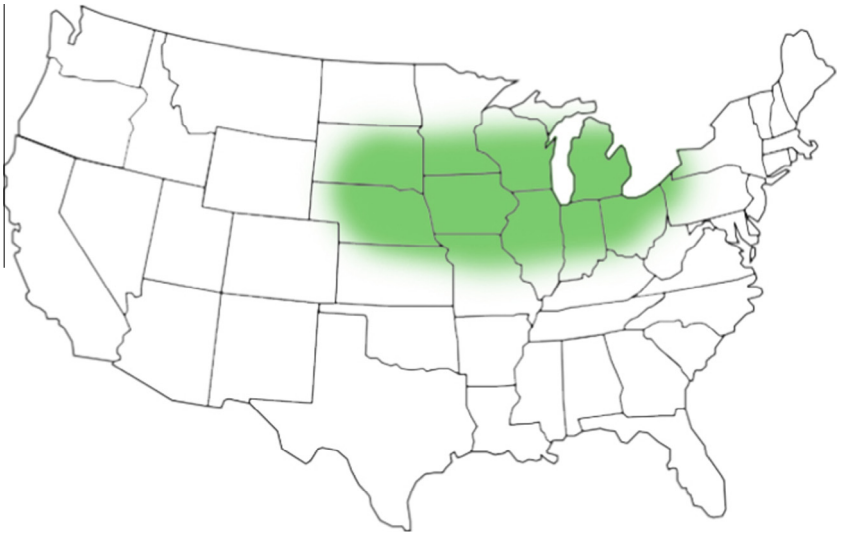

Fig. 1. Shaded is the region in the United States where farms are located.

Table 1
List of features attributed to samples provided for Syngenta Crop Challenge. Features obtained from Syngenta R&D, ISRIC, CONUS, NASS and FAO datasets are marked with numbers 1–5 respectively.

Season (year) [1]
Farm's geographic latitude[1]
Farm's geographic longitude[1]
Probability of growing soybeans in the nearby area [1,4]
Probability of field irrigation in the nearby area[5]
Probability of growing soybean of relative matureness 2.5–3[1]
How often do farmers grow soybean in the area[1]
Soil class based on texture, available water holding capacity, and soil drainage[1]
Percentage of clay in soil[2,3]
Percentage of silt in soil[2,3]
Percentage of sand in soil[2,3]
Available water capacity of soil[3]
Soil pH value[2]
Soil cation exchange capacity[2]

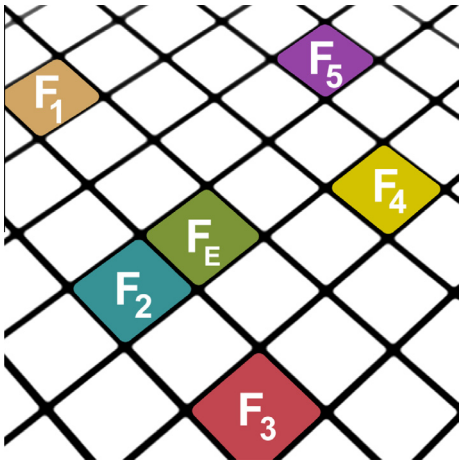

Fig. 2. Evaluation farm (F_E) and farms (F_1 to F_5) where the variety was planted, for which the yield is predicted.

the weights of the opinions, we needed a monotonically descending function of distance. We used the simplest function with such a property, which is

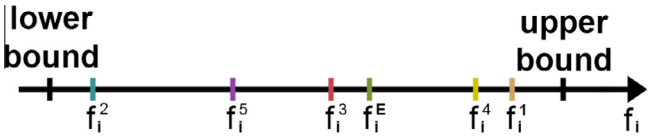

Fig. 3. Values of feature f_i on evaluation and training farms.

Table 2
Distances in feature's space between training farms and evaluation farm.

Farm	Distance
F_3	d_3
F_4	d_4
F_5	d_5
F_1	d_1
F_2	d_2

$$w = \frac{1}{1+d}. \tag{1}$$

It is actually $1/d$ shifted left for 1, so that it did not reach infinity when the distance was zero i.e. when a training farm had precisely the same value of f_i as the evaluation farm (Fig. 4). Other monotonically descending functions, such as this one with d^2 instead of d or exponential function (e^{-d}), did not give better results.

In the next step we used weights and yields to form a weighted histogram of yield at the evaluation farm for some feature f_i (Fig. 5). Unlike the classical histogram, where bins are filled with the number of farms whose yield falls within that particular range, in weighted histogram the bins are filled with weights of those farms.

In this manner we calculated the weighted histograms for all the features and all the soy varieties thus getting a 2D matrix of histograms, as in Table 3. The only difference was with the soil class feature, which was categorical. Without further information about the physical meaning of categories it was not possible to find an adequate distance measure between the categories. Therefore, we assigned the unit weights to the farms that had the same soil class as the evaluation farm, and zero weights to others. In this way only the farms with exactly the same soil class contributed to the histogram formation.

Thinking of weighted histograms of individual features as weak classifiers whose combination yields a strong one, the proposed algorithm can be interpreted within ensemble learning framework. Hence the next step in which we combine weighted histograms resembles AdaBoost algorithm (Freund and Schapire, 1997). But first, as these histograms were filled with different weights nonlinearly, we normalised them and transformed them into probability density functions – PDFs. For each of them we took expected values $E(PDF_i)$ and averaged them across all the features f_i, leaving only one (final) value per variety. This predicted value aggregates the information from all the farms on which the variety was planted, weighted according to the similarity of the whole set of features. However, not all weak classifiers were equally significant. Some were simply more accurate than the others and we had to take that into account. Each of the weak classifiers (PDFs) was assigned a coefficient c_i according to its significance, thus predicting the yield Y as

$$Y_{predicted} = \sum_{i=1}^{18} c_i * E(PDF_i), \tag{2}$$

where 18 denotes the total number of features.

Since this equation is linear, the problem of finding the optimal weak classifier weights is convex, which means that the local minimum is also the global minimum, making the optimisation pretty much straightforward (Boyd and Vandenberghe, 2004). We used CVX modelling framework for convex optimisation (Grant et al., 2014; Grant and Boyd, 2008) and tried two of its variants – constrained and unconstrained. In constrained case, the weights were limited to non-negative values, while in unconstrained variant, there were no limits whatsoever. The logic behind the constrained variant was that the equation can be viewed as the weighted average and this statistical method uses weights greater or equal to 0. On the other hand, the problem can be viewed as an unconstrained optimisation problem of finding the linear combination that minimises the error. It is defined only by the objective function that needs to be minimised (root mean square error of prediction) and does not contain neither equality nor inequality constraints (Boyd and Vandenberghe, 2004). In this case, the weights can have negative values, which are assigned to features that tend to overestimate the yield. In the end, with both variants, we took the expected value of the final PDF as the predicted yield of that particular variety for that particular farm.

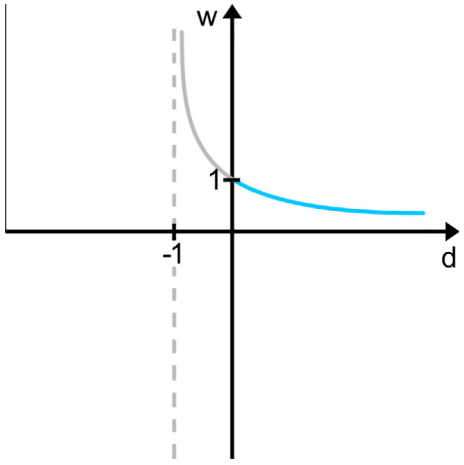

Fig. 4. Training farm's weight as a function of distance to the evaluation farm in feature space.

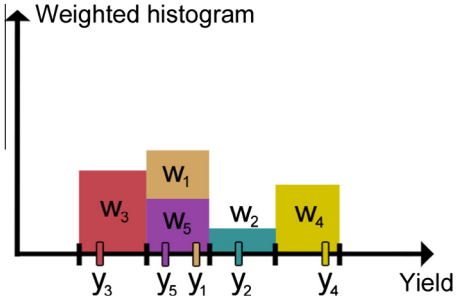

Fig. 5. Weighted histogram – each bin accumulates the weights of training farms whose yield falls in the given range. Symbols y_i and w_i denote the yield at the farm i and its relative weight.

3.2. Choosing the right varieties

Surely we could have proposed the variety that had the highest predicted yield to be planted on the whole evaluation farm, but relying on a single seed variety would have been very risky. Aiming for a slightly lower yield that came with a much lower risk was a far better alternative. In order to achieve that we reached for the portfolio optimisation theory. The necessary input consisted of:

1. Predicted yield of each seed variety.
2. Variance of yield for each seed variety.
3. Covariance between the yields of different seed varieties.

We calculated variance and covariance in the following way. The model was trained on data from 2008 to 2013 and yield was predicted for each farm from 2014 and each variety (Table 4).

Random variable corresponding to seed variety is denoted with a capital letter V. We calculated the covariance between every pair of varieties (V_i and V_j) as the covariance between the two random variables, whose realisations were known.

$$c_{ij} = E[(V_i - E[V_i])(V_j - E[V_j])] \tag{3}$$

At that point we had all the necessary inputs for calculating the efficient frontier of portfolio optimisation (Fig. 6). Efficient frontier is a curve that encompasses the points representing all the portfolios that are Pareto-optimal, meaning that there are no portfolios with the same risk that have a better yield, nor portfolios with a lower risk for that particular yield.

Choosing the right portfolio is a trade-off between yield and risk (Markowitz, 1952). Aiming for high yield could be very risky and aiming for low risk could bring poor yield. To find the optimal point on the efficient frontier, we had to set a portfolio yield threshold (PYT, dashed line on Fig. 7). For each farm in 2014 we chose the portfolio with the lowest risk whose cumulative yield did not fall below this threshold.

In this particular example, the portfolio encompassed the underlined varieties in Fig. 7. These varieties are also shown in Table 5 along with their portions in the portfolio.

The other threshold we had to set was the variety occurrence threshold (VOT). We had to discard varieties that occurred too few times to successfully approximate their PDFs. To estimate the optimal value, we varied this threshold and analysed the results.

Table 3
Matrix of weighted histograms for the evaluation farm. Rows represent different features (f_i) and columns different soybean varieties (v_i).

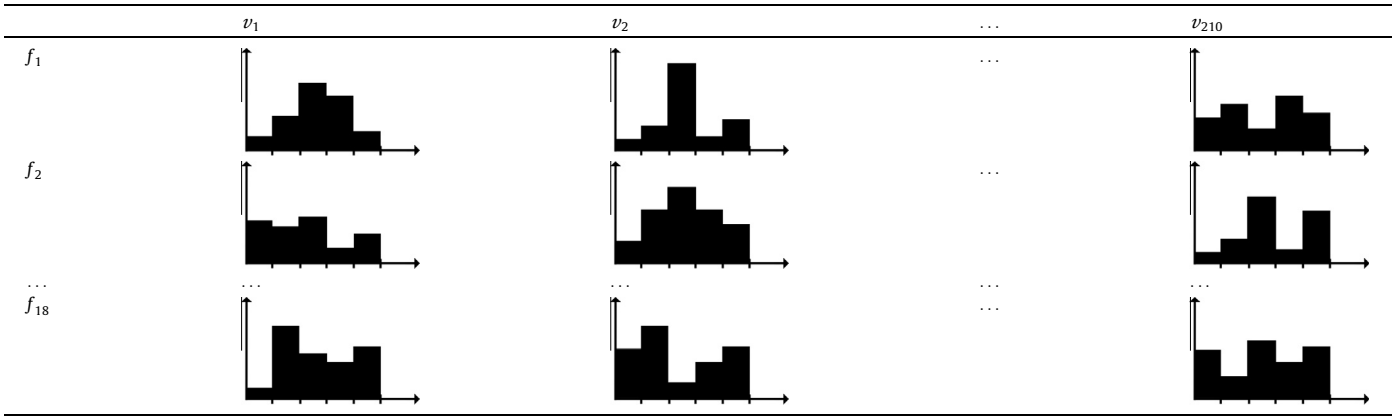

Table 4
Predicted yield (Y) for each variety and each farm from season 2014. The first index corresponds to farm number, the second to variety number.

	V_1	V_2	...	V_{210}
F_1	$Y_{1,1}$	$Y_{1,2}$	...	$Y_{1,210}$
F_2	$Y_{2,1}$	$Y_{2,2}$	...	$Y_{2,210}$
...	...	...	...	...
F_{70}	$Y_{70,1}$	$Y_{70,2}$	...	$Y_{70,210}$

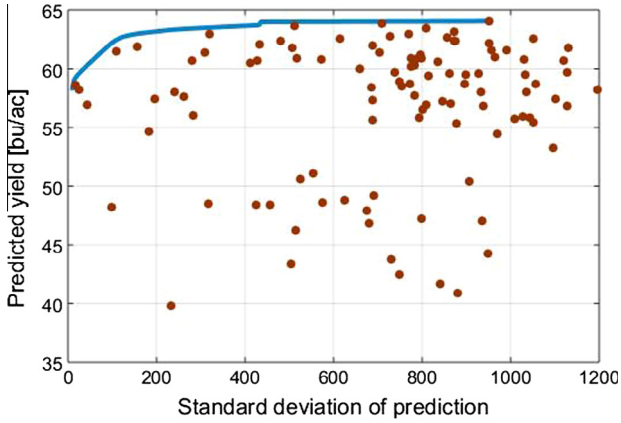

Fig. 6. Efficient frontier of portfolio optimisation (curve above the circles). Each circle represents a different seed variety.

Finally, the third parameter we needed to optimise was the covariation matrix multiplier (CMM). It is estimated that in order to have a reliable covariance matrix, number of observations per variable needs to be at least one order of magnitude higher than the number of variables (Ledoit and Wolf, 2003). In this case, there were between 100 and 200 variables, depending on the variety occurrence threshold, with around 150 observations on average. Since the number of observations was insufficient, the covariance matrix showed higher dependencies. One way to overcome this is to multiply the matrix by a constant, effectively increasing the variance and covariance of the varieties.

In order to find the right values of these three parameters, we set up the optimisation problem. Improvement I at the farm i was calculated as

$$I_i = \frac{Y_{portfolio_i} - \overline{Y_{real_i}}}{\overline{Y_{real_i}}}, \quad (4)$$

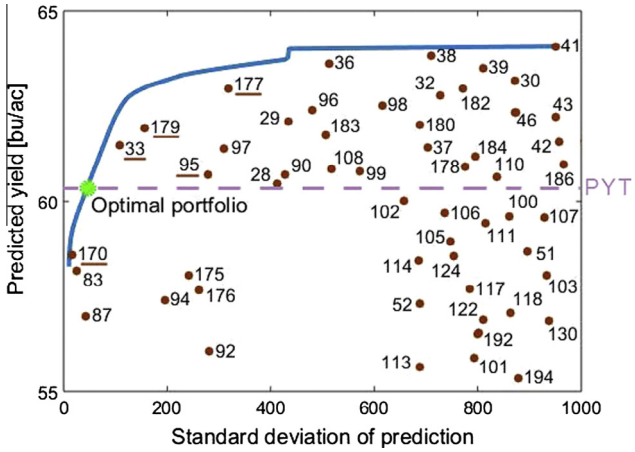

Fig. 7. Optimal portfolio consisted of the underlined varieties.

Table 5
Constituents of the optimal portfolio.

Variety	Percentage (%)
v_{33}	10.97
v_{95}	10.98
v_{170}	45.98
v_{177}	14.31
v_{179}	17.76

where $Y_{portfolio_i}$ is the yield of the proposed portfolio and $\overline{Y_{real_i}}$ is the average yield at that farm. The cost function is defined as the average improvement across all farms. The only constraint to the optimisation problem was that, according to Syngenta Crop Challenge rules, there could be up to five varieties planted on one farm and none of them could cover less than 10% of the land. The optimum was then found using grid search. All of the possible combinations were tried out and the one with the lowest cost function was selected as optimum.

4. Results

We compared proposed WHR method with different state-of-the-art algorithms in Weka, machine learning and data mining software (Hall et al., 2009). Some of them, particularly k-NN, proved to be very accurate at describing the model i.e. showed very small errors with cross-validation across the whole training set (2008–2014). The problem was that this measure of accuracy left

the possibility of training samples being used for prediction of the values within the same year, which is an impossible scenario in practice. We therefore needed to test the problem in a more realistic manner. The classifiers were trained on 2008–2013 data and tested on the data from 2014. We tested three variants of WHR:

1. equ-WHR: equal (not optimised) weights.
2. con-WHR: optimised and constrained weights.
3. unc-WHR: optimised and unconstrained weights.

As the relevant measures of accuracy we took root mean squared error, mean absolute error and correlation coefficient. Proposed method proved to be the best one according to all criteria, as shown in Table 6.

Generally, the problem with cross-validation is that there is always a risk of overfitting or underfitting. For this reason, the algorithm was both 2- and 10-fold cross-validated and both constrained and unconstrained variants were tested in the process of variety selection.

We used grid search to find optimal values of variety occurrence threshold (VOT), portfolio yield threshold (PYT) and covariance matrix multiplier (CMM). We took the values for the first two linearly with steps 10 and 0.0025, respectively. Taking the quadratic nature of variance into account, we chose exponentially growing values for the matrix multiplier – $10^0, 10^1, \ldots, 10^4$. Optimal sets of parameters and the values of cost function are shown in Table 7. The results were compared to the case where portfolio optimisation is not used, i.e. only the soy variety with the highest predicted yield is chosen.

Mean improvement is visualised with respect to variability of each individual parameter, while other parameters were left at the optimal point achieved for con-WHR (10-fold). In Fig. 8 we can see that lower values of VOT decreased the improvement, because they allowed yield PDFs to be modelled based on few samples. Higher values, on the other hand, allowed for good modelling, but narrowed the choice of seeds, thus lowering the diversity of varieties available for portfolio optimisation.

Setting the PYT lower than optimum would have meant that we would be choosing portfolios with a lower risk, but with a low yield as well. Yet, setting the threshold too close to 100% would have meant that we would be choosing only the most promising seed variety, with perhaps a small portion of other ones, which would be very risky and decreased yield in the long term (Fig. 9). The far right point (100%) shows essentially the value of improvement without portfolio optimisation (0.31%).

Fig. 10 illustrates the influence of CMM. Too low CMM left too small values in the covariance matrix, making us too confident in the predicted yields. With such a covariance matrix we would choose only the most promising variety with perhaps a small portion of others that lower the insecurity. Too high CMM resulted in too big portion of these "backup" varieties taking the focus off the most promising one.

The statistics for test farms from 2014, are shown in Table 8. We see that our method brought improvement to more than 80% of farms. There were ones where the portfolio yield was below average, but on the other hand, there were farms with the yield improved for as much as 23%. In Fig. 11, we showed the difference between average yields on the test farms and yields of the portfolios.

Table 6
Comparison of different regression methods.

Classification method	RMSE	Mean absolute error	Correlation coefficient (%)
k-NN (Cleary and Trigg, 1995)	14.674	11.863	20.26
Linear regression	11.201	9.005	26.09
Additive regression (Friedman, 2002)	11.056	8.742	28.15
Regression by discretisation	12.652	10.172	24.97
ANN (multilayer perceptron)	29.159	23.050	11.00
REPTree	13.508	11.035	19.49
equ-WHR	10.446	8.374	20.36
con-WHR (2-fold CV)	10.198	8.216	20.86
con-WHR (10-fold CV)	10.256	8.405	16.33
unc-WHR (2-fold CV)	**9.313**	**7.365**	**41.71**
unc-WHR (10-fold CV)	9.342	7.485	33.38

Table 7
Comparison of different WHR variants and the best result without the use of portfolio optimisation.

WHR variant	VOT	PYT	CMM	Mean improvement (%)
equ-WHR	110	0.995	1000	4.46
con-WHR (2-fold)	100	0.9975	1000	4.34
con-WHR (10-fold)	90	0.995	1000	4.87
unc-WHR (2-fold)	100	0.9975	10	3.40
unc-WHR (10-fold)	90	0.9925	100	3.46
The best result without the use of portfolio optimisation				0.31

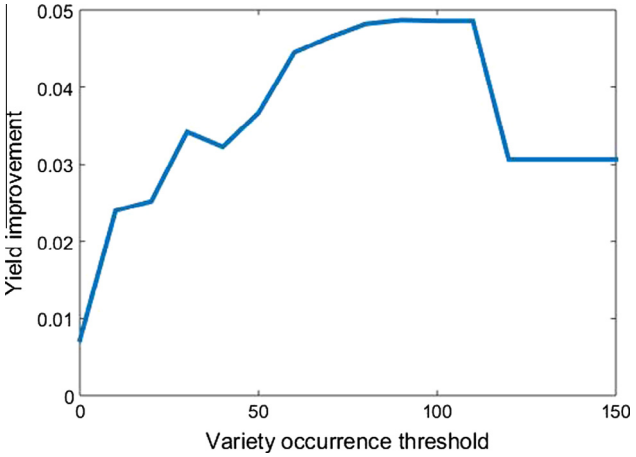

Fig. 8. Variety occurrence threshold vs. average yield improvement on test farms.

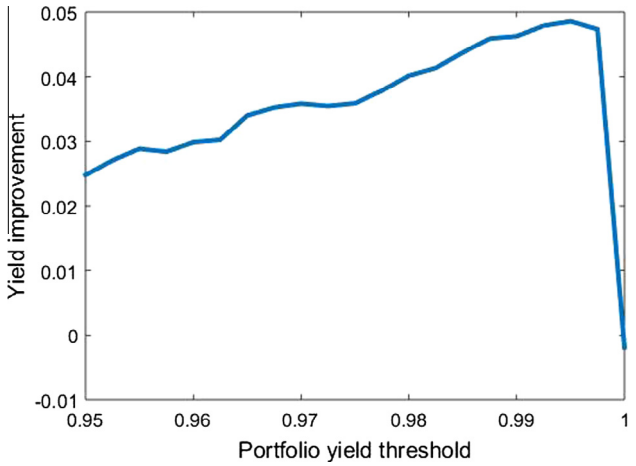

Fig. 9. Portfolio yield threshold vs. yield improvement.

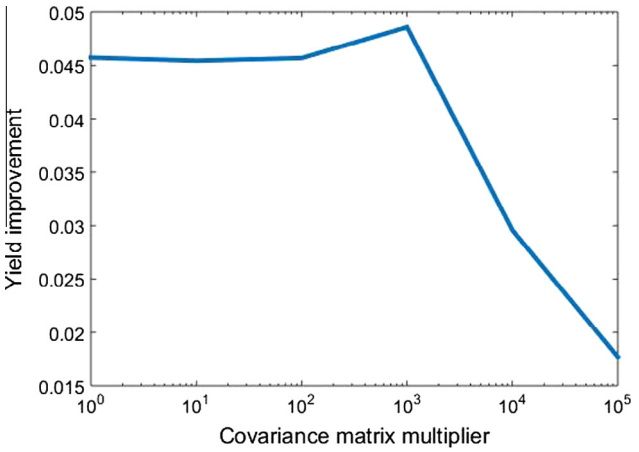

Fig. 10. Covariance matrix multiplier vs. average yield improvement on test farms.

Table 8
Statistics for test farms.

Biggest improvement	23.02%
Biggest decrease	−8.18%
Mean improvement	4.87%
Percentage of farms which witnessed improvement	82.86%

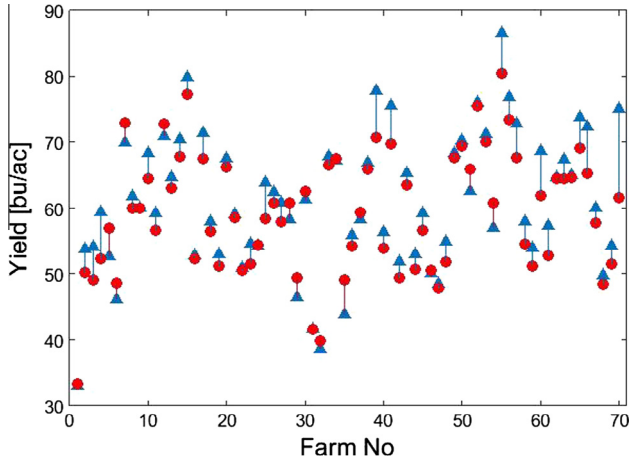

Fig. 11. Portfolio yield (triangles) vs. average yield at test farms (circles).

5. Discussion and conclusion

We showed that WHR is a very useful tool which in the case of yield prediction performed far better than conventional algorithms. The reason is that it does not discard any information in contrast to k-NN and it is more controllable than ANN, allowing for further optimisation to take place. As for the regression trees, the problem is that they are nondeterministic polynomial-time complete (NP-complete) (Hyafil and Rivest, 1976). In practice, greedy algorithms are used, but they are only locally optimal. Also the advantage of WHR over ANNs and regression trees is that it is much faster. It is interesting that although unc-WHR had lower error than con-WHR, con-WHR proved to be a better option. The reason is that having no constraints allowed unc-WHR to overfit the data, which is always a possibility with optimisation of regression and classification algorithms and insufficiently large datasets.

What we also showed is that when it comes to seed selection, using portfolio optimisation increases the yield improvement for more than 15 times, comparing to the case where only the single most promising variety is chosen. This proves what we intuitively known, that diversifying the investment, i.e. spreading the risk over a few seed varieties, is a wise strategy. The extreme case of diversification is planting all available seed varieties on a farm. It includes both those varieties that are suitable for the particular weather and soil conditions and those that are not and these extras and losses in yield even out. Although not feasible in practice, such a strategy minimises the risk and it is exactly what we compared our portfolio to. The fact that our portfolio selection strategy outperformed the minimal-risk solution confirms its effectiveness and high potential for practical use.

It is important to mention that not all seed varieties were planted at all the test farms. Often, the varieties we recognised as the most promising were not planted. We could not consider them in the portfolio optimisation, because we were not able to calculate the true yield of that portfolio. We thus believe that our results would be even better if we had more data, especially for the soy varieties which were planted on few farms.

Generally in agriculture, achieving high yield is not a problem, but it comes at a price of high investments in irrigation, pesticides and mechanisation. The biggest advantage of the method proposed in this paper is that we proved that the yield can be increased in a reliable way with no additional costs whatsoever. Therefore, we believe that our work was a valuable contribution to Syngenta Crop Challenge and is applicable for any other practical case where historical data is available.

References

Aggelopoulou, A., Bochtis, D., Fountas, S., Swain, K.C., Gemtos, T., Nanos, G., 2011. Yield prediction in apple orchards based on image processing. Precis. Agric. 12, 448–456.

Ash, C., Jasny, B.R., Malakoff, D.A., Sugden, A.M., 2010. Feeding the future. Science 327 (pp. 797–797).

Barkley, A., Hawana Peterson, H., Shroyer, J., 2010. Wheat variety selection to maximize returns and minimize risk: an application of portfolio theory. J. Agric. Appl. Econ. 42, 39.

Boryan, C., Yang, Z., Mueller, R., Craig, M., 2011. Monitoring US agriculture: the US department of agriculture, national agricultural statistics service, cropland data layer program. Geocarto Int. 26, 341–358.

Boyd, S., Vandenberghe, L., 2004. Convex Optimization. Cambridge University Press.

Brdar, S., Ćulibrk, D., Marinković, B., Crnobarac, J., Crnojević, V., 2011. Support vector machines with features contribution analysis for agricultural yield prediction. In: Proceedings of the Second International Workshop on Sensing Technologies in Agriculture, Forestry and Environment (EcoSense 2011), Belgrade, Serbia, April, 2011, pp. 43–47.

Cleary, J.G., Trigg, L.E., 1995. K*: an instance-based learner using an entropic distance measure. Proceedings of the 12th International Conference on Machine Learning, vol. 5, pp. 108–114.

Comaniciu, D., Ramesh, V., Meer, P., 2003. Kernel-based object tracking. IEEE Trans. Pattern Anal. Mach. Intell. 25, 564–577.

Di Falco, S., 2012. On the value of agricultural biodiversity. Annu. Rev. Resour. Econ. 4, 207–223.

Dragicevic, A., Lobianco, A., Leblois, A., 2016. Forest planning and productivity-risk trade-off through the Markowitz mean-variance model. For. Policy Econ. 64, 25–34.

Drummond, S.T., Sudduth, K.A., Joshi, A., Birrell, S.J., Kitchen, N.R., 2003. Statistical and neural methods for site-specific yield prediction. Trans. ASAE 46, 5.

FAO, 2016, AQUASTAT Website, Food and Agriculture Organization of the United Nations (FAO). <http://www.fao.org/nr/water/aquastat/irrigationmap/USA/index.stm> (Accessed on 27th of June 2016).

Freitas, F.D., De Souza, A.F., de Almeida, A.R., 2009. Prediction-based portfolio optimization model using neural networks. Neurocomputing 72, 2155–2170.

Freund, Y., Schapire, R.E., 1997. A decision-theoretic generalization of on-line learning and an application to boosting. J. Comput. Syst. Sci. 55, 119–139.

Friedman, J.H., 2002. Stochastic gradient boosting. Comput. Stat. Data Anal. 38, 367–378.

Gassner, A., Coe, R., Sinclair, F., 2013. Improving food security through increasing the precision of agricultural development. Precis. Agric. Sustain. Environ. Protect., 34

Godfray, H.C.J., Beddington, J.R., Crute, I.R., Haddad, L., Lawrence, D., Muir, J.F., Pretty, J., Robinson, S., Thomas, S.M., Toulmin, C., 2010. Food security: the challenge of feeding 9 billion people. Science 327, 812–818.

Gonzalez-Sanchez, A., Frausto-Solis, J., Ojeda-Bustamante, W., 2014. Predictive ability of machine learning methods for massive crop yield prediction. Span. J. Agric. Res. 12, 313–328.

Grant, M., Boyd, S., 2008. Graph implementations for nonsmooth convex programs. In: Blondel, V., Boyd, S., Kimura, H. (Eds.), Recent Advances in Learning and Control, Lecture Notes in Control and Information Sciences. Springer-Verlag Limited, pp. 95–110 http://stanford.edu/boyd/graph_dcp.html.

Grant, M., Boyd, S., 2014. CVX: Matlab Software for Disciplined Convex Programming, Version 2.1. <http://cvxr.com/cvx>.

Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., Witten, I.H., 2009. The WEKA data mining software: an update. SIGKDD Explorat. 11.

Hanson, C., 2013. Food Security, Inclusive Growth, Sustainability, and the Post-2015 Development Agenda, Background Research Paper submitted to the High Level Panel on the Post-2015 Development Agenda. United Nations, New York.

Hengl, T., de Jesus, J.M., MacMillan, R.A., Batjes, N.H., Heuvelink, G.B., Ribeiro, E., Samuel-Rosa, A., Kempen, B., Leenaars, J.G., Walsh, M.G., Gonzalez, M.R., 2014. Soil grids 1 km – global soil information based on automated mapping. PLoS One 9, e105992.

Hyafil, L., Rivest, R.L., 1976. Constructing optimal binary decision trees is np-complete. Inform. Process. Lett. 5, 15–17.

Jia, W., Zhang, H., He, X., Wu, Q., 2006. Gaussian weighted histogram intersection for license plate classification. 18th International Conference on Pattern Recognition, 2006, ICPR 2006, vol. 3. IEEE, pp. 574–577.

Kaul, M., Hill, R.L., Walthall, C., 2005. Artificial neural networks for corn and soybean yield prediction. Agric. Syst. 85, 1–18.

Ledoit, O., Wolf, M., 2003. Improved estimation of the covariance matrix of stock returns with an application to portfolio selection. J. Empir. Finan. 10, 603–621.

Liakos, V., Tagarakis, A., Fountas, S., Nanos, G., Tsiropoulos, Z., Gemtos, T., 2015. Use of NDVI to Predict Yield Variability in a Commercial Apple Orchard. In: Precision Agriculture'15. Wageningen Academic Publishers, pp. 188–197.

Louette, D., Smale, M., 2000. Farmers' seed selection practices and traditional maize varieties in Cuzalapa, Mexico. Euphytica 113, 25–41.

Marinković, B., Crnobarac, J., Brdar, S., Antić, B., Jaćimović, G., Crnojević, V., 2009. Data mining approach for predictive modeling of agricultural yield data. In: Proceedings of the First International Workshop on Sensing Technologies in Agriculture, Forestry and Environment (BioSense09), Novi Sad, Serbia, October, 2009, pp. 1–5.

Markowitz, H., 1952. Portfolio selection. J. Finan. 7, 77–91.

McGuire, S., Sperling, L., 2016. Seed systems smallholder farmers use. Food Secur. 8, 179–195.

Miller, D.A., White, R.A., 1998. A conterminous United States multilayer soil characteristics dataset for regional climate and hydrology modeling. Earth Interact. 2, 1–26.

Nalley, L.L., Barkley, A.P., 2010. Using portfolio theory to enhance wheat yield stability in low-income nations: an application in the Yaqui valley of northwestern Mexico. J. Agric. Resour. Econ., 334–347

Nalley, L.L., Barkley, A., Watkins, B., Hignight, J., 2009. Enhancing farm profitability through portfolio analysis: the case of spatial rice variety selection. J. Agric. Appl. Econ. 41, 641–652.

Pantazi, X., Moshou, D., Alexandridis, T., Whetton, R., Mouazen, A., 2016. Wheat yield prediction using machine learning and advanced sensing techniques. Comput. Electron. Agric. 121, 57–65.

Papageorgiou, E., Aggelopoulou, K., Gemtos, T., Nanos, G., 2013. Yield prediction in apples using fuzzy cognitive map learning approach. Comput. Electron. Agric. 91, 19–29.

Paydar, Z., Qureshi, M.E., 2012. Irrigation water management in uncertain conditions application of modern portfolio theory. Agric. Water Manage. 115, 47–54.

Romero, J.R., Roncallo, P.F., Akkiraju, P.C., Ponzoni, I., Echenique, V.C., Carballido, J.A., 2013. Using classification algorithms for predicting durum wheat yield in the province of Buenos Aires. Comput. Electron. Agric. 96, 173–179.

Saruta, K., Hirai, Y., Tanaka, K., Inoue, E., Okayasu, T., Mitsuoka, M., 2013. Predictive models for yield and protein content of brown rice using support vector machine. Comput. Electron. Agric. 99, 93–100.

Sperling, L., Boettiger, S., Barker, I., 2014. Integrating seed systems. Plan. Scale Brief 3.

Syngenta Crop Challenge. <https://www.ideaconnection.com/syngenta-crop-challenge/> (Accessed on 21st June 2016).

Uno, Y., Prasher, S., Lacroix, R., Goel, P., Karimi, Y., Viau, A., Patel, R., 2005. Artificial neural networks to predict corn yield from compact airborne spectrographic imager data. Comput. Electron. Agric. 47, 149–161.

Weng, Y., Crowe, K., Parker, W., Lindgren, D., Fullarton, M., Tosh, K., 2013. Using portfolio theory to improve yield and reduce risk in black spruce family reforestation. Silv. Genet. 62, 232–238.

Yengoh, G.T., 2012. Determinants of yield differences in small-scale food crop farming systems in Cameroon. Agric. Food Secur. 1, 1.