# APPLICATION OF DATA ANALYTICS TECHNIQUES TO OPTIMIZE AGRICULTURAL YIELDS

*Shokoufeh Mirzaei, College of Engineering, California State Polytechnic University, 3801 West Temple Avenue, Pomona, CA 91768, 909-869-2411, smirzaei@csupomona.edu*

## ABSTARCT

This paper provides a method which helps farmers make seed variety decisions that reliably reduce risk and increase yield. To this end, we use the dataset provided by a leading agricultural company- Syngenta Agrochemical Company - regarding the historical data of a target site for planting soybeans. First, given the yield of soybean varieties in other sites and by forecasting the growing condition, we will predict the yield of soybean varieties in the site of interest. Then, given the predicted growing conditions and the expected yield of soybean varieties, we develop an optimization formulation which provides the best mix of varieties that maximize the yield in the evaluation site with an acceptable level of risk for the farmer.

*Key words:* optimization, support vector machine, regression, machine learning

## INTRODUCTION

As world population increases needs for food and consequently land devoted to agriculture increases. However, the land dedicated to agriculture keeps declining. In addition to land, agriculture has always been highly dependent on climate pattern variations; because solar radiation, temperature, and precipitation are the main drivers of crop growth. Since the environmental variables cannot be controlled by farmers, one solution to the increasing demand with the declining land resources is to optimize the productivity of available agricultural land. This goal can be achieved by optimum allocation of lands to crop varieties subject to the weather risk and resource constraints. Therefore, to optimize a land yield, one of the major challenges that farmers frequently face is to decide what variety of products to plant each season. Specifically, this paper is a response to the challenge proposed by a company who is a leader in agriculture across the globe. The proposed challenge is associated with soybean varieties in which farmers need to decide what combination of soybean varieties each year needs to be planted so that the annual yield is maximized. However, the farmer does not any have prior information regarding the soybean yield in his own farm/land. Nonetheless, he has information regarding the soil and weather condition of other locations and their soybean yields. To tackle this complicated problem we propose a 3-step solution method that helps farmers making insightful decisions regarding the portfolio of crops for an upcoming season:

**1)** First, given the historical data, we predict the growing condition of the evaluation site for the upcoming season.

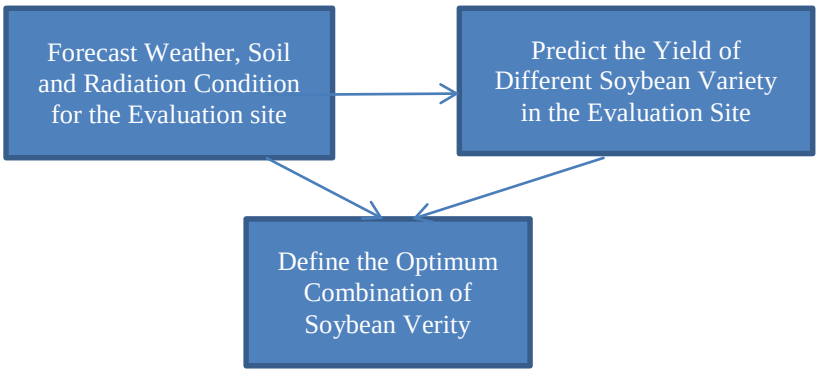
**2)** Second, we propose predictive models using regression and support vector machines to predict the expected yield of different soybean varieties for the evaluation site.

**3)** Third, by using the expected growing condition of the evaluation site (from step 1) and the expected yield of soybean varieties in the evaluation site (from step 2), we develop a linear programming problem which gives the best mix of soybean types for an optimistic, pessimistic and average scenarios. This way farmer can decide the plan which fits his/her risk profile the most.

Since majority of data analytical methods –e.g. regression or support vector machine— and their parameters are data-driven quantitative results are explained as we explain the method in section 3. In the following section an overview of the seed varieties selection method is provided.

## CRITERIA USED TO SELECT THE SEED VARIETIES

In our approach there is no one set of criteria that directly choose seed varieties. Rather, we propose a sequence of data analytics techniques for seed selection. Figure 1 shows a schematic of the proposed method. We first predict the weather, soil and radiation condition of the evaluation site. The output of this step is the input to our predictive model in the next step, in which we use machine learning technique (e.g. support vector machine and regression) to predicts the yield of soy bean varieties for the evaluation site. The output of this predictive model and the result of first step together will be used to develop an optimization model which maximizes the yields by considering a farmer's behavior toward risk. In the next section, our method is explained in details.




.

Figure 1 pathway of the proposed method

## METHODOLOGY AND QUANTITATIVE RESULTS

In this section, we provide the details of methods used for the optimization of soy bean yields in the evaluation site. Section 3.1 explains the method of forecasting used to predict the growing condition of the evaluation site. Section 3.2 explains how we use this result to predict the yield of soybean varieties in the evaluation site. In section 3.3, we will develop a linear programming which maximizes yield subject to farmers' behavior toward risk.

## DEVELOPING FORECASTING MODEL FOR WEATHER CONDITION

In this section we are mainly interested to develop a method by which we can forecast the growing condition of the evaluation site. We are specifically interested in forecasting the value of variable SEASON1, given in the training site, for the evaluation site. This is especially important for two reasons: 1) this variable indicates the crops growing condition; 2) it is available for all the experimental sites in our training set. The variable SEASON1 --which is provided in the training dataset-- is a categorical variable that "describe soil characteristics (texture, awc & drainage) and weather (season)/climate characteristics (temperature, precipitation and radiation)". Therefore, this variable contains majority of information regarding the soil, weather and radiation condition which are crucial for crop growth and therefore for planning the upcoming season. At the time of decision making, farmers do not know the exact value of variable SEAON1 for the next season. However,

using historical data --which is often publicly available-- it is possible to provide insight regarding this decision variable.
To predict the value of SEAON1 for the evaluation site we used the dataset of daily weather condition at the evaluation site provided by the company. By using this dataset, we predicted the average value of minimum temperature, maximum temperature, solar radiation, wind speed, pressure and participation of the evaluation site (i.e. variables dayl-s, prcp-mmday, srad-W-m-2, swe-kg-m-2, tmax-deg-c, tmin-deg-c, vp-Pa). By forecasting these values one can determines the appropriate value of the variable SEASON1 associated with the evaluation site in the next season. This variable is a combination of three scores which shows the soil characteristic, weather and climate conditions (Low (1), medium (2) and high (3)).

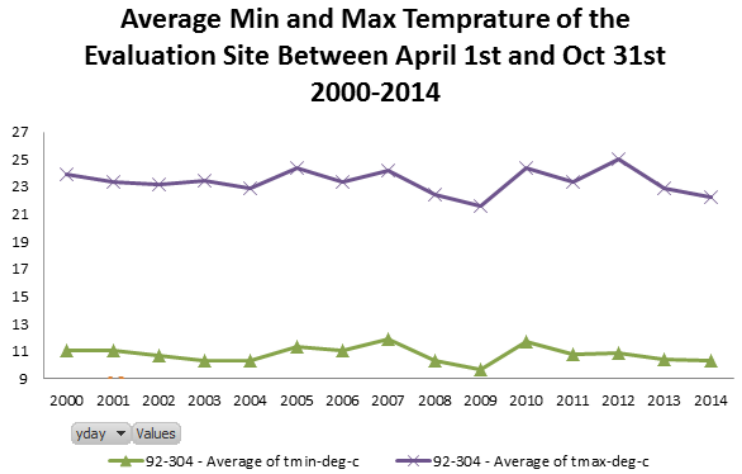



Figure 2. time series associated with the maximum and minimum temperature of the evaluation site

For forecasting the seven decision variables introduced earlier, we considered their annual average between April 1st and Oct 31st in a time series spanning from 2000 to 2014. Figure 2 shows the time series associated with the maximum and minimum temperatures of the evaluation site. We Applied the Exponential Smoothing State Space model as described in Hyndman and Khandakar [1], as it performed extremely well on the provided dataset. To implement the model we used R package *forecast* (See Hyndman and Khandakar [1]).
Figure 3 shows the decomposition of the forecasting method for the average minimum temperature between April 1$^{st}$ and Oct 31$^{st}$ for years 2000 through 2014. The three-character string in the title of the chart identifies the method used by the framework terminology of Hyndman and Khandakar [1]. The first letter indicates the error type, the second letter represents the trend type, and the third letter shows the season type. In this case, the first letter --“A"=additive-- shows the error is additive. Also, since we don’t have trend and seasonality in the dataset, the second and third characters are “N”=none. The ETS( "ANN") is in fact the simple exponential smoothing with additive errors. We used the same method for all the other 6 variables. Table 1 shows the prediction error using common comparison metrics in the literature. In the table, ME stands for the mean of errors, RMSE stands for the root mean squared of errors, MAE indicates the mean of absolute errors, MPE is the mean of percent errors, and MAPE is the mean of absolute percent errors. The result shows the selected forecasting method performs well across different evaluation criteria.

Using the EST (A,N,N) method we predicted the value of 7 decision variable for the evaluation site. Table 2 shows the forecasted values of decision variables for the upcoming season in the evaluation site with two level of confidence 80% and 95%. Given these values, one can determine the class of SEASON1 variable for the evaluation site in the upcoming season.

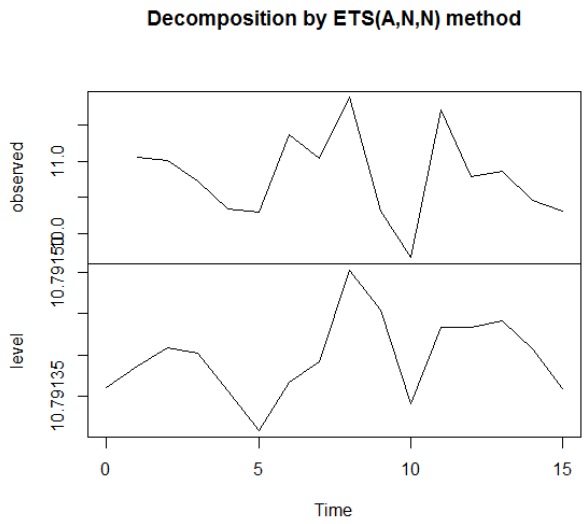



Figure 3 The decomposition of the forecasting method for the average minimum temperature between April 1st and Oct 31st for years 2000 through 2014

Table 1. Forecasting error using different metrics suggested in the literature

| Row Labels | ME | RMSE | MAE | MPE | MAPE |
|---|---|---|---|---|---|
| "Average of t-min.deg.c" | 0.001 | 0.572 | 0.458 | 0.286 | 4.247 |
| "Average of t-max.deg.c" | 0.000 | 0.879 | 0.680 | 0.143 | 2.920 |
| "Average of vp.Pa" | 0.016 | 57.737 | 47.637 | 0.171 | 3.434 |
| "Average of swe.kg.m.2" | 0.000 | 0.881 | 0.498 | -inf | Inf |
| "Average of srad.W.m.2" | 0.000 | 4.131 | 3.107 | -0.013 | 0.860 |
| "Averageofprcp.mmday" | 0.167 | 0.800 | 0.650 | 1.098 | 18.119 |

Table 2. Forecasted values of growing condition variables for the upcoming season in the evaluation site

| **Average value between April 1st and Oct 31st** | **Point Forecast** | **Lo 80** | **Hi 80** | **Lo 95** | **Hi 95** |
|---|---|---|---|---|---|
| tmin.deg.c | 10.79136 | 10.05878 | 11.52394 | 9.670977 | 11.91174 |
| tmax.deg.c | 23.37584 | 22.24973 | 24.50194 | 21.65361 | 25.09806 |
| vp.Pa | 1384.46 | 1310.467 | 1458.454 | 1271.298 | 1497.623 |
| swe.kg.m.2 | 0.302949 | -0.82672 | 1.432616 | -1.42473 | 2.030625 |
| srad.W.m.2 | 362.4729 | 357.1795 | 367.7664 | 354.3773 | 370.5686 |
| prcp.mmday | 3.922567 | 2.808784 | 5.03635 | 2.219183 | 5.625952 |

Lo 80 and Lo 95 are the lower bound of 80% and 95% confidence intervals, respectively. Also, Hi 80 and Hi 95 are the upper bound of 80% and 95% confidence intervals, respectively.

Suppose the chosen level of confidence for this problem is 95%. Then, we can define three alternatives for the variable SEASON1 using the two boundary values as well as the point estimation. The boundary conditions will represent the extreme cases for SEAON1 – optimistic and pessimistic. Consequently, we can use these three SEASON1 values to predict the yield of soy bean verities at the evaluation site under boundary conditions as well as the average condition.

## PREDICTING THE YIELD OF EVALUATION SITE GIVEN THE WEATHER CONDITION OF THE EVALUATION SITE

Although, historical data regarding the performance of different soybean varieties might not be available for the site in hand, the performance of soybean varieties in different locations and with different weather and soil condition might provide a good indication of the yield of soybean verities in the site of interest. To this end, we study the training data set provided using two alternative methods to predict the yield of soybean varieties in the given site.

## DATASET AND FEATURES

The company has provided a dataset consisting of the results of different experiments conducted in different sites to evaluate the yield of growing soybean varieties. This dataset is imbalanced in the sense that the number of experiments conducted in different sites is not the same and often only one experiment is represented for a given soybean variety, breeding ,weather and soil conditions. To tackle this problem, for each location, weather condition, season, soybean variety and breeding, the average yield of the experiments is considered as the yield reprehensive of the soybean variety. This reduces the size of original dataset from 34212 rows to 9959 rows. In the new dataset, multiple experiments conducted in a site with similar control conditions are aggregated and represented by their average. This way, we make sure for a given location and controlled conditions, soybean variety is represented at most once in the dataset.

## PREDICTING THE YIELD OF SOY BEAN VARIETIES USING REGRESSION

In this section, we first investigate the characteristics of the curated dataset we created. In Figure 4, the graph in the left shows the box-plot of the average yield in our curated dataset. As it is shown in Figure 1, there are values in the dataset which are outside of the whiskers of the box-plot. Therefore, it is safe to remove them from the dataset. Additionally, in Figure 4, the graph in the right shows the histogram of the yields after removing the 83 outliers. The figure shows, regardless of the land location and growing condition, the average of soybean yields follow a normal distribution with average of 59.64 and standard deviation of 10.81.

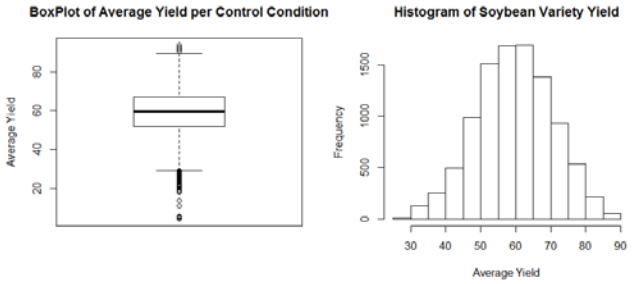



Figure 4 the box-plot whisker and histogram of the average yields of soybean variety in the training set

This distribution of data in the figure 4 justifies the use of regression for predicting the yield of soybean verity with different growing condition. To implement a linear multiple regression, dummy variables were introduced for categorical features. Three variables are, SEAON1, RM-BAND, CLIMATE, BREEDING_G and VARIERT according to the definition of variables provided in the dataset. Also, in this study SEASON is considered as a trend decision variables to capture the impact of time if necessary.

In this part, first, we develop a regression model to predict the soybean variety given the input parameters associated with location and growing condition of the site. To this end, we implemented an iterative elimination approach to remove the insignificant factors from the pool of decision variables in the training dataset.

In the problem statement provided, it is not clear how the growing condition determines the SEASON1 variable. Therefore, we assume given the boundary conditions of table 2 the pessimistic condition is presented by 111 and optimistic scenario is presented by 333 and 222 is the average case. With these assumptions and using the regression model developed in this section, we calculated the expected yield of soybean varieties at the evaluation site for three scenarios mentioned earlier. Figure 5 shows the graph related to the optimistic prediction of yields for the evaluation site sorted from the largest to the smallest value.

## PREDICTING THE YIELD OF SOY BEAN VARIETIES USING SUPPORT VECTOR MACHINE

In this section, we adopt a Support Vector Machine regression (SVM-e) –as an alternative to the regression method provided in section 3.2.2--to predict the yield of different varieties and breeding for the site of interest by using the 3 predicted growing conditions (SEANS1).

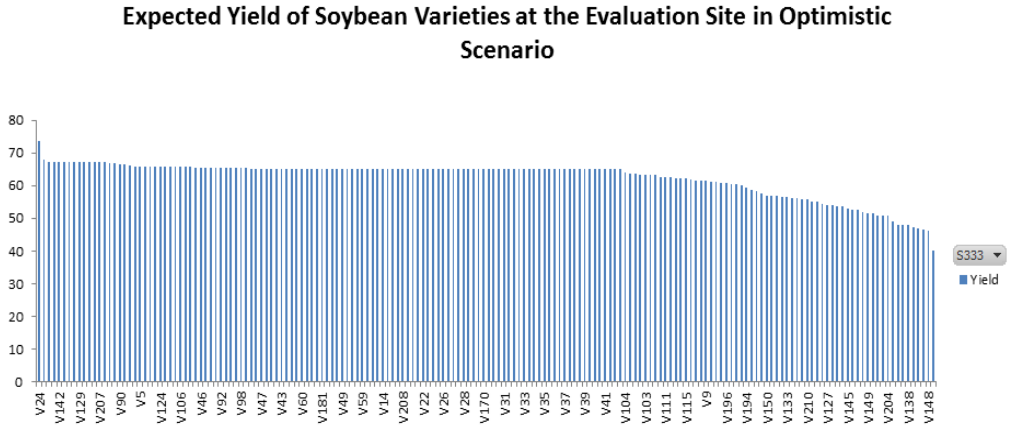



Figure 5 optimistic yields of soybean varieties in the evaluation site sorted from the largest to the smallest

To this end, VARIETY_YI, is our response variable which is a continuous decision variable. SVM regression is based on ideas drawn from statistical learning theory [2]. In this case, the support vectors are defined so that the error $z_i = wx_i + b - y_i$ is within an interval $[-\varepsilon, \varepsilon]$ i.e. $z_i \in [-\varepsilon, \varepsilon]$. However, to allow for non-linearity of the model the condition will be enforced in the objective function rather than being enforced by rigid constraint:

$$\min \quad \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{n} \max\{0, 1 - y_i(wx_i + b)\} \quad (1)$$

In this study, Gaussian Radial Kernel provided the most accurate result in predicting the GDT_TS. This Kernel function is provided in equation 2.

$$f(z, z') = e^{y\|z - z'\|^2} \quad (2)$$

where $\|z - z'\|$ is the squared Euclidean distance between the two feature vectors and gamma is a parameter. The gamma in equation 2 and the cost parameter c in equation 1 are the parameters of the SVM-e which need to be determined respectively.

From the company training data, two subsets were generated to train and test the predictive model. Our dataset includes 50 features and we used a cross validation method to validate the model. In each round of the training process, we selected 80% of the data to train and 20% of the data to test the result. The selection of training and test set were implemented randomly and with replacement. Along with the training process a grid search was conducted to tune the SVM-e parameters γ, c, and ε. To this end, during each round of validation, a combination of parameter was used to train and test the model. The process was repeated 120 times with $\gamma = \{0.00001, 0.0001, 0.001, 0.01, 0.1, 1, 10, 20, 30, 40\}$ and $c = \{0.1, 0.2, 0.3, 0.4, 0..5, 1, 2, 3, 4, 5, 6, 7\}$, and $\varepsilon = 0.1$. Also, we found that a Radial Basis Kernel Function provides the lowest error term [3]. Figure 2 shows the average of root mean square error (MRMSE) of predicted yields for each combination of gamma and cost parameters and for the test sets. The final parameter values are γ=0.01, c=3, and ε=0.1. The parameters were selected so that the errors of both training and test sets are reasonably low. We used these selected parameters and the associated model to predict the yield of soybean variety at the evaluation site. Note that one of the input variables for developing the predictive model was SEASON1, which we calculated in section 2. By comparing the two predictive models --SVM and regression-- we concluded that regression is a more effective method of prediction for the given dataset. It is mainly because the parameters found for SVM are local optimums. By conducting an exhaustive search, it is likely to find a more effective set of parameters

for the SVM which could lead to better results than regression. However, due to the time constraints, we will use the results of regression to predict the soybean yields in the evaluation site. Finally, by comparing the results from the optimistic, pessimistic and average scenarios we notices that five soybean varieties consistently provide the top 5 yields in the evaluation site. These soybean varieties are are V24, V96, V205, V 159 and V142. In the next section we will explain how we select an optimum mix of the varieties.

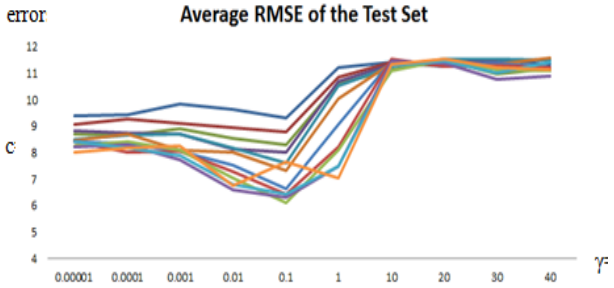



Figure 6 RMSE for each combination of gamma and cost for the test sets.

## OPTIMIZATION OF THE YIELD OF SOY BEAN VARIETIES

In this section, we provide a simple mathematical programming which will provide the optimum mix of soybean varieties for an upcoming season. We solve the optimization problem for the optimistic, pessimistic and average scenarios.

In addition to the risk imposed by weather and growing conditions, it is important to note that the amount of land dedicated to a specific variety also imposes a risk to the problem. For example, allocating 100% of the land to Variety 1 is a more risky decision than allocating it to two different types of soybeans. We denote the risk function associated to allocation of lands to varieties by function $f(x) = ax^2$ where $x$ is the portion of land allocated to a variety and $a$ is a constant. In this function, as the amount of land allocated to a product increases the risk associated to it increase quadratically. Therefore, this function avoids the allocation of land to only one variety i.e. putting all eggs in one basket! In contrast, it will distribute the land among varieties. Given the risk term, in the following we define the mathematical modeling used to allocate lands to varieties.

Notations:

$X_i$ percent of land allocated to plant variety i, i=1,2,…,182

$L$ maximum amount of land available to the farmer

$E_i$ expected yield per acre predicted for the site of interest (this is the result of prediction from previous step)

$Y_i$ binary variable identifying whether we plant variety i or not

$$Max \quad \sum_{i=1}^{182} L x_i y_i - \sum_{i=1}^{182} (x_i * 10)\text{^}2 \qquad (3)$$

$$\sum_{i=1}^{182} x_i = 1 \qquad (4)$$

$$\sum_{i=1}^{182} y_i \leq 5 \qquad (5)$$

$$x_i \geq 0.10 - M(1 - y_i) \qquad \forall i \in \{1,2 \dots ,182\} \qquad (6)$$

$$x_i \leq y_i \qquad \forall i \in \{1,2 \dots ,182\} \qquad (7)$$

$$x_i \geq 0 \, , \, y_i \in \{0,1\} \qquad (8)$$

Equation 3 is the objective function which consists of two terms; the first term is the total yield and the second term is the risk associated with the allocation of land to varieties. Equation 4 enforces the summation of land percentage allocated to soybean varieties becomes 1. Inequality 5 makes sure that not more than 5 varieties are allocated to the evaluation site. Equations 6 and 7 together make sure that when a land allocated to a variety it is at least 10% of the land. Term 8 determines the range of our decision variables.

Although the mathematical programming provided above is nonlinear, linearization of the nonlinear term is straight forward. We solved the proposed optimization problem for the problem in hand using the risk parameter of a=20. By implementing the optimization model for optimistic, pessimistic, and average yields, the following unanimous result was obtained from the three scenarios.

$V_{24}=0.69$ $V_{96}=0.11$ $V_{205}=0.10$ $V_{159}=0.10$

## CONCLUSION

As world population and need for food increases, the optimum usage of the land dedicated to agriculture becoming more critical. In addition to land, agriculture has always been highly dependent on climate pattern variations including solar radiation, temperature, and precipitation, which are important in crop growth. In this paper, given the information regarding soy varieties and the land that they have been planted we needed to predict the varieties in a new land. To this end, we provided a three-step approach to determine how a farmer can make seed variety decisions that reliably reduce risk and increase yield. Using predictive methods and optimization techniques we provided a hybrid method which can suggest the best mix of varieties that maximize the yield in the evaluation site with an acceptable level of risk for farmers. This approach can further be improved by embedding methods that projects the farmers' utility function and thus their level of risk.

## ACKNOWLEDGEMENT

The author would like to acknowledge Syngenta Agrochemical Company for providing the problem statement and dataset regarding this problem through Syngenta Crop Challenge.

## REFERENCES

[1] R. Hyndman, Y. Khandakar, Automatic time series forecasting: the forecast package for R 7, 2008, URL http://www. jstatsoft. org/v27/i03, (2007).
[2] J. Weston, A. Gammerman, M. Stitson, V. Vapnik, V. Vovk, C. Watkins, 1 Support Vector Density Estimation, (1999).
[3] B. Schölkopf, K. Tsuda, J.-P. Vert, Kernel methods in computational biology, MIT press, 2004.