Computer program for construction of regression function for phenotype in agroclimatic models with interactions

K.N. Kozlov[1]*, S.V. Nuzhdin[1, 2], M.G. Samsonova[1]

[1] *Peter the Great St. Petersburg Polytechnic University, St. Petersburg, Russia*
[2] *University of Southern California, Los Angeles, CA, USA*

DOI 10.18699/ICG-PlantGen2019-58

* e-mail: kozlov_kn@spbstu.ru

Abstract: Regression models that connect agronomic traits to climatic factors provide valuable insights into phenological characteristics of cultivars. The genotype-by-environment interactions are modeled by a weighted sum of pairwise products between a control functions and group indicator variables. In contrast to existing modeling frameworks in our approach the analytic form of a control function, regression coefficients and a set of predictors are inferred by stochastic minimization of the deviation of the model output from data. The approach was successfully applied to the three datasets for soybean and chickpea to predict time to flowering with coefficient of determination 0.45-0.97.

Key words: mathematical modeling; regression; agroclimatic factors.

1. Introduction

Plants react to climate change with changing phenotype integrating climate-biospheric interactions (Morisette et al., 2009). Though mathematical modeling is one of the most important tools for prediction of phenologiclal traits the accuracy remains a problem (Richardson et al., 2012). The duration of developmental stages must closely coincide with the available season for acceptable results. Widely used methods for prediction of phenological traits are calculation of the sum of temperatures above the temperature minimum and regression models (Major et al., 1975; Pedersen et al., 2004; Setiyono et al., 2007). Several successful crop simulation models like SSM (Soltani et al., 2006a, 2006b), DSSAT (Boote et al., 2013; Jones et al., 2003, 2017a), APSIM (Keating et al., 2003) and others (Battisti et al., 2018; Williams et al., 1989) have been developed for legumes. Biophysical and biochemical processes are described with differential equations problem-specific parameters for genotype, soil, weather and economic factors. Developed in the absence of genomic information, these models considered genotype influence at best as a set of given "genetic coefficients" that do not correspond to actual genes (Hwang et al., 2017). Consequently, the inability of these models to take gene-by-environment interactions into account restricts the prediction of phenological traits of cultivars across different geographical locations and genotypes (Vadez et al., 2013). We propose a more general approach implemented in the computer program called 'nlreg', in which the analytic form of a control function together with regression coefficients and a set of predictors are inferred automatically by stochastic minimization of the deviation of the model output from data.

2. Materials and methods

The interactions between factors and *K* different geographical locations or genotypes is modeled by a weighted sum of pairwise products between the control functions F_n and the group indicator variables d_i^k such that d_i^k=1, for plant i from group k and =0 otherwise. Thus, for a set of data records (y_i, X_i), where y_i is the phenotype and X_i is the vector of climatic factors for plant i, the computer program presented here constructs Model (1).

$$y_i = \beta_0 + \sum_{n=0}^{N-1} \beta_{n+1} F_n(X_i) + \sum_{n=0}^{N-1} \sum_{l=1}^{K} \zeta_{k \cdot N+n} \cdot F_n(X_i) \cdot d_i^k + \varepsilon_i \quad (1),$$

where β_n and $\zeta_{k \cdot N+n}$ are the regression coefficients, N is the number of functions F_n, and ε_i is the standard error.

The analytic form of function F_n is constructed from the vector of codons of length M using Grammatical Evolution (GE) (Noorian et al., 2016; O'Neill and Ryan, 2001) which utilizes a context-free grammar (CFG). The CFG is defined by the 4-tuple of a finite set of terminal symbols, non-terminal symbols, the production rule set and the start symbol (Aho et al., 2006). In our approach, non-terminal symbols are defined as arithmetic operations "+", "–", "*", "/" or expressions *X*, (*X* – *Const*), or 1/(*X* – *Const*) where the members of terminals set *X* and *Const* denote a name of the predictor and the constants, respectively.

The model is further built using the LASSO algorithm (Tibshirani, 1996) which minimizes the sum of squared differences between model output and data and penalizes the sum of absolute values of regression coefficients β_n, thus reducing non-important ones to 0.

The vector of codons is determined by minimizing the approximation error using the stochastic optimization technique called the Differential Evolution Entirely Parallel (DEEP) method (Kozlov and Samsonov, 2011; Kozlov et al., 2016). Differential Evolution (DE) was proposed in 1995 (Storn, 1995, 1997). DEEP incorporates several recent enhancements (Fan and Lampinen, 2003; Kozlov et al., 2016; Zaharie, 2002). DEEP employs the pool of worker threads with an asynchronous queue of tasks to evaluate the individual solutions in parallel. The code is available on GitLab (https://gitlab.com/mackoel/deepmethod).

Although a few GE implementations are freely available (Noorian et al., 2016; Peter Harrington, 2018), they either lack a specific set of expressions or show low performance in our tasks. We implemented GE in C++ using Armadillo

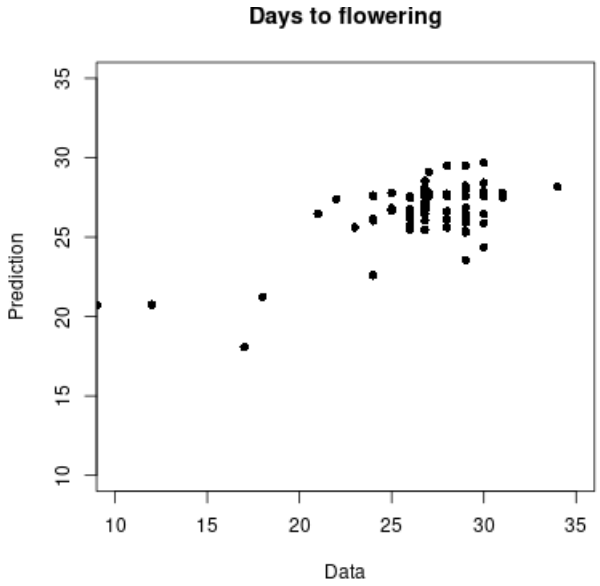

Figure 1. Comparison of model predictions with experimental data for models for chickpea VIR landraces from Turkey and Ethiopia.

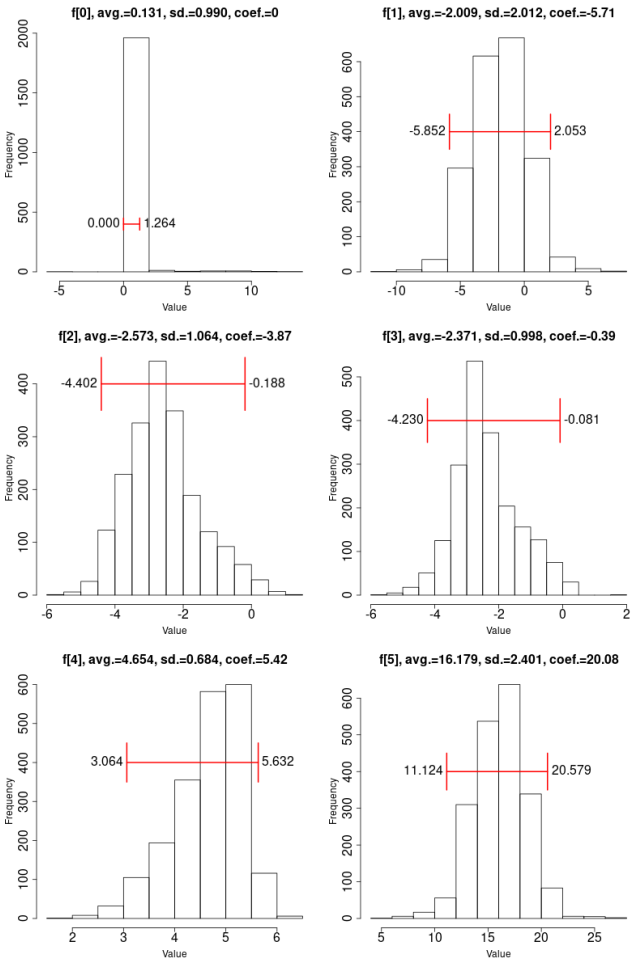

Figure 2. Confidence intervals for the 95% significance level for the intercept and 5 regression coefficients for functions F_n.

(Sanderson and Curtin, 2016), mlpack (Curtin et al., 2013), HDF5, HighFive (The Blue Brain Project, 2018) and Qt for efficient matrix operations, the LASSO method, data input-output and utility functions, respectively. The code is available on GitLab (https://gitlab.com/mackoel/nlreg) and can be compiled for GNU\Linux or MS Windows 8.1 and 10 operating systems.

The program is accessed using a command line interface that accepts several options. Tabular data is read from a HDF5 file and parameters are supplied in a file in INI-format. To facilitate high-performance computing, the program can utilize OpenMP and MPI parallelization technologies.

3. Results and discussion

The approach was successfully applied to the three datasets for soybean and chickpea to predict time to flowering. For a dataset that comprises 379 plants of 9 different soybean accessions phenotyped at Pushkin VIR stations in 1999-2013, the method constructed a more accurate model (coefficient of determination $R^2 = 0.60$) than the previous one in (Kozlov et al., 2018; Seferova and Novikova, 2015).

The models for chickpea VIR landraces from Turkey ($R^2 = 0.45$) and Ethiopia ($R^2 = 0.52$) were built in (Kozlov et al., 2019b). The comparison of model predictions with experimental data is presented in Figure 1. Modeling revealed the difference in the impacts of temperature and precipitation. The impact of temperature was 60 and 48 % for Turkey and Ethiopia, respectively. The impact of precipitation was estimated at 86 and 89 % for Turkey and Ethiopia, respectively.

The model for wild chickpea collected by von Wettberg et al. (2018) ($R^2 = 0.97$) showed that the genotype-by-environment interactions accounted for about 17.2 % of variation in time to flowering (Kozlov et al., 2019a).

To access the practical identifiability of model parameters, we applied a bootstrap approach (Mudelsee, 2010) and performed 1999 runs with sampled datasets (Efron and Tibshirani, 1993). Confidence intervals for the 95 % significance level for the intercept and 5 regression coefficients for functions F_n are presented in Figure 2. Five out of six coefficients are considered identifiable as their confidence intervals do not contain zeroes. Genotype-by-environment interactions were significantly non-zero ($P < 0.05$ in t-test) for 56 out 90 combinations of functions F_n with allele combinations at 6 SNP positions.

4. Conclusions

In contrast to existing modeling frameworks, in our approach control functions are automatically composed in analytic form that allows a wider range of non-linear dependencies between the phenotype and climatic factors to be explored. The results of numerical experiments with the wild chickpea dataset showed that certain environmental variables differently affect the flowering time of different genotypes. The analysis revealed that the 95 % confidence intervals for five out of six regression coefficients did not contain zeroes and thus represent a well-established influence of the climatic factor on time to flowering. 56 regression coefficients of genotype-by-environment interactions are significantly non-zero. Consequently, the computer program developed is a

useful tool for mathematical modeling of phenological traits like flowering time and the investigation of genotype-by-environment interactions.

References

Aho A.V., Lam M.S., Sethi R., Ullman J.D. Compilers: Principles, Techniques, and Tools (2Nd Edition). Boston, MA, USA: Addison-Wesley Longman Publishing Co., Inc., 2006.

Battisti R., Sentelhas P.C., Boote K.J. Sensitivity and requirement of improvements of four soybean crop simulation models for climate change studies in Southern Brazil. *International Journal of Biometeorology.* 2018;62:823–832.

Boote K.J., Jones J.W., White J.W., Asseng S., Lizaso J.I. Putting Mechanisms into Crop Production Models. *Plant Cell Environ*. 2013.

Curtin R.R., Cline J.R., Slagle N.P., March W.B., Ram P., Mehta N.A., Gray A.G. mlpack: A Scalable C++ Machine Learning Library. *Journal of Machine Learning Research.* 2013;14:801–805.

Efron B., Tibshirani R. An introduction to the bootstrap. New York: Chapman & Hall, 1993.

Fan H.-Y., Lampinen J. A Trigonometric Mutation Operation to Differential Evolution. *Journal of Global Optimization*. 2003;27:25.

Hwang C., Correll M.J., Gezan S.A., Zhang L., Bhakta M.S., Vallejos C.E., Boote K.J., Clavijo-Michelangeli J.A., Jones J.W. Next generation crop models: A modular approach to model early vegetative and reproductive development of the common bean (*Phaseolus vulgaris* L). *Agricultural Systems*. 2017;155:225–239.

Jones J.W., Hoogenboom G., Porter C.H., Boote K.J., Batchelor W.D., Hunt L.A., Wilkens P.W., Singh U., Gijsman A.J., Ritchie J.T. The DSSAT cropping system model. *European Journal of Agronomy*. 2003;18:235–265.

Jones J.W., Antle J.M., Basso B., Boote K.J., Conant R.T., Foster I., Godfray H.C.J., Herrero M., Howitt R.E., Janssen S. et al. Brief history of agricultural systems modeling. *Agricultural Systems*. 2017a; 155:240–254.

Keating B., Carberry P.S., Hammer G., Probert M.E., Robertson M.J., Holzworth D., Huth N.I., Hargreaves J., Meinke H., Hochman Z. et al. An overview of APSIM, a model designed for farming systems simulation. *European Journal of Agronomy.* 2003;18:267–288.

Kozlov K., Samsonov A. DEEP -- Differential Evolution Entirely Parallel Method for Gene Regulatory Networks. *Journal of Supercomputing.* 2011;57:172–178.

Kozlov K., Samsonov A.M., Samsonova M. A software for parameter optimization with Differential Evolution Entirely Parallel method. *PeerJ Computer Science*. 2016;2:e74.

Kozlov K., Singh A., Berger J., Wettberg E.B., Kahraman A., Aydogan A., Cook D., Nuzhdin S., Samsonova M. Non-linear regression models for time to flowering in wild chickpea combine genetic and climatic factors. *BMC Plant Biology.* 2019a;19:94.

Kozlov K.N., Novikova L.Yu., Seferova I.V., Samsonova M.G. A Mathematical Model of the Effect of Climatic Factors on Soybean Development. *Biophysics.* 2018;63:136–137.

Kozlov K.N., Samsonova M.G., Nuzhdin S.V. Regression Model For Time To Flowering Of Chickpea Landraces. *Russian Journal of Genetics.* 2019b;55:1–5.

Major D.J., Johnson D.R., Tanner J.W., Anderson I.C. Effects of daylength and temperature on soybean development. *Crop Science.* 1975;15:174–179.

Morisette J.T., Richardson A.D., Knapp A.K., Fisher J.I., Graham E.A., Abatzoglou J., Wilson B.E., Breshears D.D., Henebry G.M., Hanes J.M. et al. Tracking the rhythm of the seasons in the face of global change: phenological research in the 21st century. *Frontiers in Ecology and the Environment.* 2009;7:253–260.

Mudelsee M. Climate time series analysis: classical statistical and bootstrap methods. Dordrecht; New York: Springer, 2010.

Noorian F., de Silva A.M., Leong P.H.W. gramEvol : Grammatical Evolution in *R*. *Journal of Statistical Software.* 2016;71:1–26.

O'Neill M., Ryan C. Grammatical evolution. *IEEE Transactions on Evolutionary Computation.* 2001;5:349–358.

Pedersen P., Boote K.J., Jones J.W., Lauer J.G. (). Modifying the CROPGRO-Soybean Model to Improve Predictions for the Upper Midwest. *AGRONOMY JOURNAL*. 2004;96:556–564.

Harrington P. Genetic Programming C++ Code. 2018.

Richardson A.D., Anderson R.S., Arain M.A., Barr A.G., Bohrer G., Chen G., Chen J.M., Ciais P., Davis K.J., Desai A.R. et al. Terrestrial biosphere models need better representation of vegetation phenology: results from the North American Carbon Program Site Synthesis. *Global Change Biology.* 2012;18:566–584.

Sanderson C., Curtin R. Armadillo: a template-based C++ library for linear algebra. *Journal of Open Source Software.* 2016;1:26.

Seferova I.V., Novikova L.Yu. Climatic factors that impact the early-maturing soybean accessions in North-West Russia. *Works on Applied Botany, Genetics and Breeding.* 2015;176:88–97.

Setiyono T.D., Weiss A., Specht J., Bastidas A.M., Cassman K.G., Dobermann A. Understanding and modeling the effect of temperature and daylength on soybean phenology under high-yield conditions. *Field Crops Research.* 2007;100:257–271.

Soltani A., Robertson M.J., Mohammad-Nejad Y., Rahemi-Karizaki A. Modeling chickpea growth and development: Leaf production and senescence. *Field Crops Research*. 2006a;99:14–23.

Soltani A., Hammer G.L., Torabi B., Robertson M.J., Zeinali E. Modeling chickpea growth and development: Phenological development. *Field Crops Research.* 2006b;99;1–13.

Storn R. Differential evolution - a simple and efficient adaptive scheme for global optimization. 1995.

Storn R. Differential Evolution – A Simple and Efficient Heuristic for Global Optimization over Continuous Spaces. *DIFFERENTIAL EVOLUTION*. 1997:19.

Tibshirani R. Regression shrinkage and selection via the lasso. *J. of the Royal Statistical Society : Series B*. 1996;58:267–288.

Vadez V., Soltani A., Sinclair T.R. Crop simulation analysis of phenological adaptation of chickpea to different latitudes of India. *Field Crops Research*. 2013;146:1–9.

von Wettberg E.J.B., Chang P.L., Başdemir F., Carrasquila-Garcia N., Korbu L.B., Moenga S.M., Bedada G., Greenlon A., Moriuchi K.S., Singh V. et al. Ecology and genomics of an important crop wild relative as a prelude to agricultural innovation. *Nature Communications*. 2018;9.

Williams J.R., Jones C.A., Kiniry J.R., Spanel D.A. The EPIC Crop Growth Model. *TRANSACTIONS of the ASAE*. 1989;32:497–511.

Zaharie D. Parameter Adaptation in Differential Evolution by Controlling the Population Diversity. In Proc. of 4th InternationalWorkshop on Symbolic and Numeric Algorithms for Scientific Computing. Petcu D. (Ed.). Timisoara, Romania: Analele Universitatii Timisoara, 2002. pp. 385–397.

Acknowledgements. The work is supported by the Federal Targeted Program (Agreement No. 14.575.21.0136 from 26.09.2017, RFMEFI57517X0136). Calculations were performed in Supercomputer Center of Peter the Great St.Petersburg Polytechnic University.

Conflict of interest. The authors declare no conflict of interest.