Agris on-line Papers in Economics and Informatics

Volume XI Number 1, 2019

Data Pre-processing for Agricultural Simulations

Jan Jarolímek[1], Jan Pavlík[1], Jana Kholova[2], Swarna Ronanki[3]

[1] Department of Information Technologies, Faculty of Economics and Management, University of Life Sciences Prague, Czech Republic

[2] Department of Crops Physiology, International Crops Research Institute for Semi-Arid Tropics - System

[3] Analysis for Climate Smart Agriculture, Hyderabad, IndiaDepartment of Crop Production, ICAR - Indian Institute of Millets Research, Hyderabad, India

Abstract

The process of agricultural simulation using APSIM requires input meteorological data to be prepared in a specific format and the simulation setting file to be ready before the simulation processing starts. Because of possible time savings when conducting large number of simulations at once, it is preferable to create all the input and settings files for all the simulations beforehand and process the simulations in batches as large as possible. This article specifically deals with the data acquisition, transformation and preparation process. It also outlines initial testing and computing time estimations and discusses scheduling, parallel processing and other possible simulation optimization methods..

Keywords

APSIM, big data, data processing, yield optimization, software automation, parallel processing.

Jarolímek, J., Pavlík, J., Kholova, J. and Ronanki, S. (2019) "Data Pre-processing for Agricultural Simulations", *AGRIS on-line Papers in Economics and Informatics*, Vol. 11, No. 1, pp. 49-53. ISSN 1804-1930. DOI 10.7160/aol.2019.110105.

Introduction

With increasing processing capabilities, it is becoming possible to tackle larger research endeavours. In the area of scenario simulations, this increase in hardware power allows for broader assignments in terms of variable combination. Historically, the total amount of simulations was severely limited and required either very narrow specification of simulation parameters or usage of techniques that lowered the processing requirement at the cost of less accurate results, such as downscaling (Hewitson and Crane, 1996). Nowadays, higher hardware power can be utilized to calculate more simulations extending the limits of the usual variable spectrum. However, the multi-linear nature of growth of number of simulations based on number of options for each variable still limits the simulation process in general, so some restrictions need to be upheld regardless.

One example of simulation software that was originally designed for small scale field simulations on a single computer but has seen a resurgence as a large scale (even on a global scale) tool for simulation of agricultural production is the Agricultural Production Systems sIMulator (APSIM). This software provides important insight into challenges regarding food security, climate change adaptation and carbon trading (Holzworth et.al., 2014).

By using supporting software tools for automation and scheduling it is possible to tackle large number of simulations in APSIM by splitting the computation onto multiple machines utilizing parallel processing as shown by (Zhao, et.al., 2013). Even though hardware and software scales differently during processing (Kambadur, et.al., 2013), with a proper setup and data pre-processing it is possible to make up for the increased number of simulations. Apart from increasing the range for variables, increasing the resolution of the grid will also affect the number of simulations required, however as pointed out by (Mass, et. al., 2002) when it comes to weather forecasts, reducing the grid size beyond certain limit no longer significantly improves the quality of results.

Another issue is also the period of input weather data. Due to changes in global climate, only short-term predictions are possible (Aurbacher, et.al.,

2013), so it might be necessary to recalculate simulations on a periodical basis with newest possible data sets, in order to maintain high level of usability of results. This however introduces additional layer of scaling, so in order to ensure up-to-date knowledge based on the simulation results, measures must be taken to reduce the processing requirements of each individual set of simulations (Skoogh, et al., 2010).

The requirements for data storage also scale based on the number of simulations. However, there are possibilities to cut down the storage requirements by extracting required results during the processing from output files that have been already calculated and deleting them. But considering the processing of simulations is the most time-consuming part of the research process, deleting finished output files may be ill-advised, since they are the most "expensive" to create. Therefore, a better solution would be to search for additional storage capacities. Luckily, thanks to the rise of IoT (Internet of Things) as a source of data (Stoces et.al., 2018), most research entities have bolstered their storage hardware in recent years.

Overall, the issue of large-scale simulations, their processing requirements and optimization in general is very current topic. Many researchers are looking for solutions in various areas, whether it be utilizing cloud-based capacities (Szufel, et al., 2017), exploiting existing hardware to its maximum potential (Fujimoto, 2016) or looking for new frameworks altogether (Kirby, et.al., 2018).

Materials and methods

In order to simulate agricultural production two input files are required for APSIM. Firstly, there is the .met file which contains historical meteorological data for a given field / grid square. The required parameters are daily rates of solar radiation (radn), minimum daily temperature (tmin), maximum daily temperature (tmax) and precipitation rate (rain). Apart from these daily values the .met file must also contain pre-calculated values for annual ambient average temperature (tav) and annual amplitude in mean monthly temperature (amp).

The second input file is the .apsim file that contains settings for the simulation (irrigation rates, sowing window, sowing density, fertilization etc.) as well as data related to the given grid square (such as soil properties, characteristics for given plant genotype and so on).

The meteorological data we use are from Goddard Institute for Space Studies (GISS), which is part of National Aeronautics and Space Administration (NASA). The AgMERRA Climate Forcing Datasets (https://data.giss.nasa.gov/impacts/agmipcf/agmerra/) are free to download in an .nc4 format. The datasets are split into files per year (from 1980 to 2010) and per variable. Therefore, some pre-processing will be required to transform the data, since APSIM is expecting the data split into files per grid square containing a table with all the values for all the variables and for all the years.

The .apsim files are just .xml files using the markup language to capture all the input variables for each given simulation. These files have to be prepared based on real agricultural conditions in given area. For the purposes of multi-variable simulation, each single simulation has to be reproduced so that every possible combination of variables was represented. Considering the large-scale nature due to the high number of grid squares as well as high number of variable combinations, it is unfeasible to do this task manually.

To complete the pre-processing, both the .met and .apsim file for all the simulations must be ready. The next task is to optimize the simulation computations themselves outside of the APSIM software. Possible solutions include parallel processing, utilization of cloud based resource structure, optimizations regarding scheduling and use of additional hardware resources during their downtime. We plan to publish a separate follow-up article regarding this process at a later date.

Results and discussion

The required data conversion from .nc4 files downloaded from the NASA into .met files required by the APSIM software was achieved using a MATLAB script. The calculation of (tav) and (amp) variables can be done within MATLAB as well or using R script. However, we found that the easiest way is to first convert to .xls, do the calculation in MS Excel and then convert to .prn file, which has the same required structure as the .met, and simply change the extension.

In order to create the settings for all the simulations we have written a program using C# language that loads a single .apsim file with one simulation in it and returns an .apsim file with all the possible variations of that simulation, with all the combinations of chosen variables. In our case, it was 12600 simulations per each grid square. This batch size proved to be too high for the APSIM software, so we had to adjust the program to create

several smaller files (see below). The choice to use C# was arbitrary based on experience of the programmers in our team. Any other programming language (python, java etc.) can be chosen and will work just as fine to write a similar program / script.

Our simulations used single soil settings for all the grid squares. In cases where different soil settings are required the preprocessing depends on the form and availability of data in a given country. This will provide additional layer of preprocessing, however as shown by (Kim, et.al., 2018), this step can be also automated by writing an application specific to the soil database that will fetch the data in bulk.

Overall, the pre-processing of data did not provide any challenges in terms of software / hardware requirements; even with high number of simulations (hundreds of millions) the computation time is in the range of several minutes. The majority of input in this stage was therefore programmer labour time needed to write the scripts and programs.

The simulation processing itself will be done using the command line version of APSIM. The software has a graphical user interface (GUI) provided (see Figure 1), but it does not include any functionality that would be helpful setting up computation of large number of simulations. It is designed merely as a tool to better visualize the contents of the .apsim files and to edit values when dealing with small number of simulations at a time. The computation time of both variants (command line and GUI) is similar, but the former provides easier options for automation and scheduling using third-party tools.

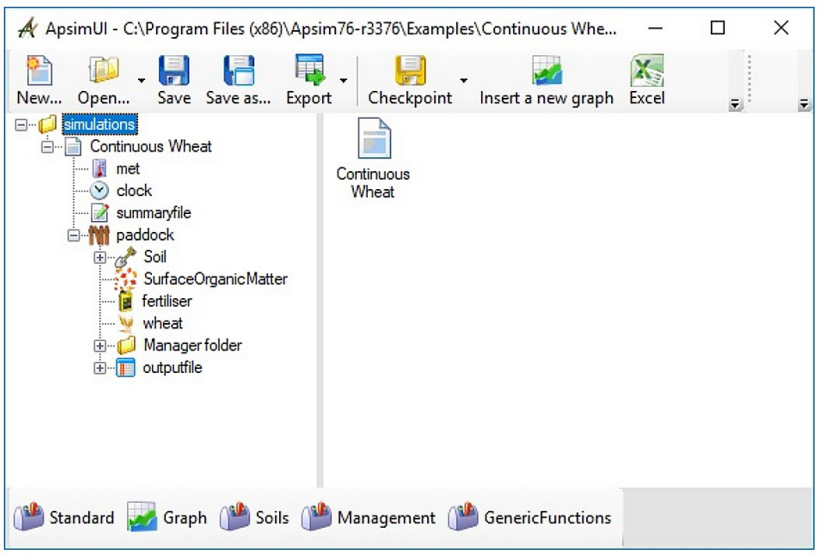
We conducted preliminary testing runs for some of the simulations on several different machines in order to estimate the overall time requirements. What we found was that the processing time doesn't scale perfectly with the amount of simulations in a single batch. Possibly due to some overhead requirements (initialization, clean-up etc.) the efficiency of simulations processed per minute goes up with the batch size (see Table 1 for approximate results).

Number of simulations	Approx. time (minutes)	Simulations per minute
100	2.5	40.0
500	11.5	43.5
1000	22.0	45.5
2000	40.0	50.0
2500	48.0	52.1

Source: own processing

Table 1: Preliminary processing efficiency for different batch sizes.

Based on these results it became clear that in order to optimize the processing, the batch size should be as large as possible. However, the APSIM software cannot handle all the simulations at once. There seem to be a limit on maximum batch size that is influenced by used hardware. Some of the stronger machines we used for testing were able to handle between 2000 and 3000 simulations at once, whereas regular desktop computers



Source: own processing

Figure 1: APSIM User Interface.

with mid-range hardware installed were not able to go over 1000 simulations in a single batch. This limit seems to be influenced by available memory capacity, but strangely during the simulation processing itself, the limiting factor was processor, not memory. This would imply that the memory capacity is mostly relevant during the initialization. We plan to conduct further testing using wider variety of hardware to reach more definite conclusion in this matter.

At this moment, the best way to optimize processing seems to be determining optimal batch size for each machine that will be involved in the computation process and use third-party scheduling software to run the simulations on every machine separately when its resources are free for use. With the way our C# program to generate simulation works at this point, that would mean creating a stockpile of simulations of varying batch sizes for each machine. Due to uneven workload of machines however, this may prove problematic, since each computer will drain its simulation stockpile at different rate. A solution to this issue might be adjusting the simulation generation so that it does not work as a static application, but rather an ongoing server application. That way the schedulers that handle processing could request batches of input files when necessary.

Conclusion

The requirements for data pre-processing when working with APSIM scale with the amount of simulations due to the lack of in-built option for variable simulations. However, this can be handled reasonably efficiently using features of MATLAB for weather data processing combined with self-written scripts to generate simulation files for all possible combinations of variables. There is little to no room for improvement or time saving when handling these necessary tasks. But when utilizing parallel processing it becomes possible to reduce computing time via optimizing the batch size for each individual machine. Having the option to select variable batch size within the simulation generation script therefore proved very advantageous.

But overall, we must conclude that the age of APSIM software really shows, especially with regards to lack of features / packages that could help with large scale research by removing or at least reducing the required pre-processing requirements. This issue is only amplified by the fact that personnel who use APSIM often do not possess enough IT knowledge and training, especially when it is required to operate additional third-party software. Similar findings regarding lack of IT expertise we pointed out by (Reinmuth and Dabbert, 2017) for instance. Some of these issues will be hopefully handled in the APSIM Next Generation as outlined by (Holzworth et. al., 2018).

Acknowledgements

This article was created with the support of the Internal Grant Agency (IGA) of FEM CULS in Prague, no. 2019A0017 „Bulk processing of large volumes of geographical data“.

Corresponding authors
Ing. Jan Pavlík
Department of Information Technologies, Faculty of Economics and Management
Czech University of Life Sciences Prague, Kamýcká 129, 165 00 Prague – Suchdol, Czech Republic
Phone: +420 224 382 356, Email: pavlikjan@pef.czu.cz

References

[1] Aurbacher, J., Parker, P. S., Sanchez, G. A. C., Steinbach, J., Reinmuth, E., Ingwersen, J. and Dabbert, S. (2013) “Influence of climate change on short term management of field crops - A modelling approach”, A*gricultural Systems*, Vol. 119, pp. 44-57. ISSN 0308-521X. DOI 10.1016/j.agsy.2013.04.005.

[2] Fujimoto, R. M. (2016) “Research Challenges in Parallel and Distributed Simulation”, *ACM Transactions On Modeling And Computer Simulation*, Vol. 26, No. 4. 10493301. DOI 10.1145/2866577.

[3] Hewitson, B. C. and Crane, R. G. (1996) “Climate downscaling: Techniques and application” *Climate Research*, Vol. 7, pp. 85-95. E-ISSN 1616-1572, ISSN 0936-577X. DOI 10.3354/cr007085.

[4] Holzworth, D., Huth, N. I., de Voil, P. G., Zurcher, E. J., Herrmann, N. I., McLean, G., ... Keating, B. A. (2014) “APSIM - Evolution towards a New Generation of Agricultural Systems Simulation.” *Environmental Modelling & Software*, Vol. 62, pp. 327-350. ISSN 1364-8152. DOI 10.1016/j.envsoft.2014.07.009.

[5] Holzworth, D., Huth, N. I., Fainges, J., Brown, H., Zurcher, E., Cichota, R., Verrall, S., Herrmann, N. I., Zheng, B. and Snow. V. (2018) “APSIM Next Generation: Overcoming Challenges in Modernising a Farming Systems Model”, *Environmental Modelling & Software*, Vol. 103, pp. 43-51. ISSN 1364-8152. DOI 10.1016/j.envsoft.2018.02.002.

[6] Kambadur, M., Tang, K., Lopez, J. and Kim, M. A. (2013) “Parallel scaling properties from a basic block view”, ACM SIGMETRICS *Performance Evaluation Review*, Vol. 41, pp. 365-366. ISSN 0163-5999. DOI 10.1145/2494232.2465748.

[7] Kim, K. S., Yoo, B. H., Shelia, V., Porter, C. H. and Hoogenboom, G. (2018) “START: A data preparation tool for crop simulation models using web-based soil databases”, *Computers and Electronics in Agriculture*, vol. 154, pp. 256-264. ISSN 0168-1699. DOI 10.1016/j.compag.2018.08.023.

[8] Kirby, A. C., Yang, Z., Mavriplis, D. J., Duque, E. P. N. and Whitlock, B. J. (2018) “Visualization and data analytics challenges of large-scale high-fidelity numerical simulations of wind energy applications”, *AIAA Aerospace Sciences Meeting*. AIAA SciTech Forum, Kissimmee, Florida. DOI 10.2514/6.2018-1171.

[9] Mass, C. F., Ovens, D., Westrick, K. and Colle, B. A. (2002) “Does increasing horizontal resolution produce more skillful forecasts? The results of two years of real-time numerical weather prediction over the Pacific northwest”, *Bulletin of the American Meteorological Society*, Vol. 83, No. 3. DOI 10.1175/1520-0477(2002)083<0407:DIHRPM>2.3.CO;2.

[10] Reinmuth, E. and Dabbert, S. (2017) “Toward more efficient model development for farming systems research - An integrative review”, *Computers And Electronics In Agriculture*, Vol. 138, pp. 29-38. ISSN 0168-1699. DOI 10.1016/j.compag.2017.04.007.

[11] Skoogh, A., Michaloski, J. and Bengtsson, N. (2010) “Towards continuously updated simulation models: Combining automated raw data collection and automated data processing”, *Proceedings - Winter Simulation Conference*, pp. 1678-1689. ISSN 08917736. DOI 10.1109/WSC.2010.5678901.

[12] Stoces, M., Masner, J., Kanska, E. and Jarolimek J. (2018) "Processing of Big Data in Internet of Things and Precision Agriculture", *Agrarian Perspectives XXVII.: Food Safety - Food Security, Proceedings of the 27th International Scientific Conference*, pp. 353-358. ISBN 978-80-213-2890-7. ISSN 1213-7979.

[13] Szufel, P., Czupryna, M. and Kaminski, B. (2017) "Optimal execution of large scale simulations in the cloud. The case of route-To-pa sim online preference simulation“, *Proceedings - Winter Simulation Conference*, pp. 3702-3703. DOI 10.1109/WSC.2016.7822408.

[14] Zhao, G., Bryan, B. A., King, D., Luo, Z., Wang, E., Bende-Michl, U., Song, X. and Yu, Q. (2013) “Large-scale, high-resolution agricultural systems modeling using a hybrid approach combining grid computing and parallel processing”, *Environmental Modelling & Software*, Vol. 41, pp. 231-238. ISSN 1364-8152. DOI 10.1016/j.envsoft.2012.08.007.