Skoltech
Skolkovo Institute of Science and Technology

Thesis Changes Log

Name of Candidate: Dmitry Shadrin

PhD Program: Computational and Data Science and Engineering

Title of Thesis: Data-driven modeling of plant growth dynamics in controlled environments

Supervisor: Prof. Maxim Fedorov

The thesis document includes the following changes in answer to the external review process.

Dear Reviewers,
I would like to thank you for the useful comments which were invaluable in improving the thesis. Below please find the responses.

Reviewer: Nikolay Brilliantov

1) Is it justified to apply Verhulst model to the leaves growth? What I mean is that this model assumes some maximal amount of the growing substance, while I doubt that there are any limitations (except for the seasonal changes) for the leaves growth. The author should comment on this.

Answer: Overall, Verhulst model is very popular for describing of biological systems. First of all, it should be noticed that Verhulst model was used for assessment of the dynamics of the projected leaves area. In the experiments 2D top-down images of plants were obtained. Using these images, the projection of leaves area was calculated. Growing plants broadwise have an effect on the values of the calculated projected leaves area. The projection of leaves area has limitations for the investigated plants (tomatoes, lettuce and cucumbers) because these plants are not able to grow infinitely broadwise. This effect was also observed experimentally. For the development of the models, data from the initial stages of growth was used, but it was assumed that the plant has some maximum projected leaves area. Also, the obtained experimental data has the good fit to the Vehulst model. For example, by fitting the Verhulst models' parameters to the dataset on tomato by using non-linear least square method ~13% average relative error was obtained. One of the types of the plant that was used in the experiments is MicroTina. This is the superdetermenistic type of plant and it has the limitations in the overall biomass growth.
For the industrial experiment with cucumbers, I investigated only the initial stage of growth. It can be observed from the experiment that after the initial stage of cucumbers growth (3-4 first leaves), they start growing upwards, and the projections obtained from the top-down images don't change significantly (changing of projected leaves area can occur due to the effect overlapping and this effect starts to prevail). The estimation and modeling of the biomass were done at the initial stage of growth. So, the biomass that is accumulated at the initial stages and estimated using the first 3-4 projected leaves area has its limitation.
This discussion was added to Sections 3.2 and 4.2, also literature references were added.

2) Integrating Eq. (3.2) it is assumed that the parameters are constant. At the same time in Figs. 3.11 and 3.12 these parameters are shown as functions of time. It should be stressed in the text (or the figure captures) that these quantities are not the true one, but just a current estimates. In the present version it is not clearly stated.

Answer: The suggested point was stressed in the text and in the figure captures. In general case parameters μ and S_{max} are time-dependent. The aim of Kalman filtering is to estimate the dynamics of changing of these parameters. Changing of these parameters are guided by the environmental conditions.
The captures were rewritten in the following way:
"True, estimated by Kalman filter and by the non-linear least square dynamics of maximum leaf area changing in time."
"Plants growth rate dynamics estimation from the experimental data is shown for each plant (number 1-9, except for the 5-th). The 2-nd plant showed the maximum growth rate which matches the experimental data."
"Simulated and estimated by Kalman filter dynamics of growth rate."

3) Eq. (3.18) does not make much sense. Please re-write this piece of the text the exposition of the material is very awkward here.

Answer: This misprint was fixed. Mass m was missed in the Eq. (3.18). The Eq. (3.18) was fixed. The exposition of the material was also improved in this part of Section 3.4 to make clearer the narration.

4) Are experiments in Chapter 4 the same as in the previous Chapters? Please be more specific.

Answer: The experiments that are described in Chapter 4 differ from the experiments described in Chapter 3. In Chapter 3 datasets obtained from small scale experiments, that described at the beginning of the chapter (Section 3.1) are used. In Chapter 4, modeling in each section was performed based on the data that was described at the beginning of each section. In the newly introduced Section 4.5, where methods were compared, the used datasets for comparison of the methods were specified.

5) There are many undefined variables and awkward expressions:
For instance, h (width?) and C (nutrients concentration?), (page 54) are not defined, "v" is mixed with "u" in page 54 du is "dimensionless unit"(page 56) – does not make sense

Answer: All the issues with undefined variables in Section 2.3 and across the thesis were fixed. "dimensionless unit" was changed to "space unit" and "time" was changed to "time unit". The section was revised and unclear statements were improved.

6) Incomplete figure capture for Fig. 3.9, wrong sentence after Eq. (3.2).

Answer: The capture was changed to:
"Hydroponic system showing the tomato growth rate within 1 month and1 week: germination, (b) vegetation, and (c-d) flowering lifetime." was changed to "Hydroponic system where tomato plants were growing during 1 month and 1 week: (a) germination stage, (b) vegetation stage, and (c-d) flowering."

The misprint in the sentence was fixed

7) Eq. (3.19) is to be better explained. The expression “water potential” is very ambiguous -- is the chemical potential meant? “n” and “w” there are not explained -- neither in the text nor in the Table 3.3 below.

Answer: The explanation of the Eq. (3.19) was revised and updated, also the explanation of the parameters was clarified.

Reviewer: David Macii

1) The whole thesis should be carefully checked and revised to fix a large number of minor typos, English mistakes or improper sentences. Some paragraphs and explanations are verbose and entangled. As a result, the Thesis is not always easy to follow.

Answer: The whole thesis was carefully checked and revised, all minor typos and English mistakes were fixed. The verbose and entangled paragraphs and sentences were rewritten.

2) Also, the style chosen for citations is very weird and it should be changed. The Author refers indeed to other papers using a format like “Name Surname [Year]” in the middle of sentences. However, in this way the Author continuously interrupts his writing, which is quite annoying for the reader. I recommend that the Author changes the style of all references, including Authors’ names between brackets (e.g. [John Smith 2001] or simply [Smith 2001]) and moving them to the end of the paragraph they refer to.

Answer: All references were revised. The style of references was changed to [John Smith 2001]. Also, references were reorganized and moved from the middle to the end of the sentence or paragraph in all cases where it was possible.

3) Background. Maybe some Sections (e.g. Sections 2.2) should be broken into smaller, more focused subsections as the current ones are very long and it is not so easy for the reader to extract the novelty of the proposed work out of such a large amount of information.

Answer: Section 2.2 was divided into the following subsections (paragraphs):

- Image based technologies for plant phenotyping.
- 3D imaging.
- Machine learning for modeling of growth dynamics.
- Non-destructive methods for assessment of seeds germination.
- Plant diseases detection.
- Embedded systems.
- Bottlenecks of ML and CV application.

4) Chapter 3 and 4. As a possible improvement, I think that the Author should highlight more clearly why different kinds of models are used to assess plant growth dynamics as well as the advantages and disadvantages of the investigated techniques so as to guide other researchers towards the best option for a given scenario. For instance, it is not so clear why or under what conditions the Kalman filter-based models are preferable to data-driven approaches. Maybe a final table summarizing pros and cons of different techniques (to be added in the Conclusions) could help. In general, it is not straightforward for the reader to understand quickly which techniques are preferable in different contexts and what their requirements and limitations are.

Answer: The discussion section (Section 4.5) was added. In this section, different kinds of models for performing prediction tasks were compared (RNN, DMD, Kalman filter). To compare developed data-driven techniques with hybrid model-based approaches, the developed data-driven approaches (RNNs) were trained and evaluated on the same datasets that were used in Chapter

3 for creating hybrid models. The results were presented, advantages and disadvantages, as well as the best options for a particular use case were discussed. The table, that summarized the results of comparison was added to the end of Section 4.5. The conclusions to the Sections 3.3 and 4.3 were revised and updated in order to discuss the advantages and limitations of the computer vision methods for growth dynamics and germination rate assessment by semantic and instance segmentation. Also, the conclusions to Sections 3.5 and 4.2 were revised and updated in order to show the advantages and limitations of the methods, proposed for biomass prediction using different techniques. The conclusion to the Section 4.4 was also revised and possible applications and limitations of the method for finding optimal wavebands were discussed.

4) I also found a bit strange that a Kalman filter is used not really to estimate the state of growth of the plant (which is indeed dynamic), but the parameters of the model (i.e., Smax and μ that look static or quasi static). Indeed, not surprisingly the system matrix is the identity matrix.

Answer: The main idea was to implement a Kalman filter for a fast estimation of the main growth parameters. The problem in estimating the real state of growth by using of Kalman filter is that it is necessary to have a robust model, that will describe plant growth dynamics (e.g. leaves growth). However, as it was shown in Section 2.3 these models are much more complex, have many unmeasured parameters and do not always have stable solutions, that is why it is quite difficult to implement them directly in Kalman filtering procedure for obtaining accurate estimations. In the developed approach it was shown how it is possible to apply Kalman filter by the linearization of measurement function (projected leaves area measurements). This allows to overcome the problem with the direct implementation of not robust models that describe the state of the plant growth.

5) Chapter 5. However, this part of the Thesis lacks generality, as it seems to be driven by too specific case studies. Thus, the generality of results and conclusions is questionable. More specifically, I appreciate the great work done by the Author to investigate such complex environmental issues as those mentioned above, by using machine learning techniques. However, the replicability/applicability of such techniques on a large scale in conditions and contexts different from those of the case studies presented can be hardly assessed due to the large variety and variability of parameters and factors that may affect this kind of analysis.
Maybe the Author could try to add one or more Sections where the aforementioned problems are formalized in more abstract and general terms, e.g. providing a list of the quantities that in general must be considered for monitoring, prior to focusing on the reported case studies.

Answer: The Section 5.1 (Problem statement and proposed solutions) was added to Chapter 5.The beginning of the chapter was reorganized and rewritten in order to show the general problems that will be solved by implementation of the data-driven techniques. Current simulators for plant growth dynamics assessment and yield prediction are using a set of controlled and uncontrolled environmental parameters as the input. The list of parameters that are considered for monitoring was presented. The main aim of this chapter is to propose and validate a set of universal data-driven techniques that allow to improve the accuracy of the currently used simulators. This can be achieved by improving the accuracy of the prediction of spatial distribution of uncontrolled parameters and by improving the accuracy of the assessment of the effects on plant growth dynamics caused by controlled parameters. It was proposed to use Gaussian process regression in couple with Bayesian information criteria for automatization and improving accuracy for prediction of spatial distribution of environmental parameters. This approach was validated on the use case of modeling of water quality parameters. The effects of controlled parameters on plant growth were assessed by using such machine learning techniques as support vector regression and artificial neural networks. These ML techniques were evaluated by the prediction of toxicity effects caused by inserting different amounts of fertilizer and pollution.

Reviewer: Belyaev Mikhail

Different tools and approaches are used in various projects, so a reader cannot directly compare these methods. I believe that a thesis should be a research work itself (not a collection of papers), and different parts should be interconnected.
More specifically, I'd like to see a direct comparison of three different types of methods (physical, hybrid, or data-driven based) for a couple of plant modeling problems.

Answer: The direct comparison of different methods is discussed in newly added Section 4.5. The comparison was performed for methods aimed at plant growth dynamics prediction (projection of leaves area). The same proposed in Section 4.1 RNN was trained and evaluated using the same data that was used for development of the Kalman filter and DMD algorithm for their direct comparison. The results were discussed. In the Section 4.5 it was also discussed the advantages and disadvantages of each method compared to other methods. The results were summarized in the table at the end of the section.

Reviewer: Vladimir Palyulin

1) I am concerned that the methods used could have their limitations. I understand that mostly the thesis is a proof of concept that data-driven methods solve and simplify many problems. This is shown pretty well. However, I would appreciate to see a brief analysis of applicability and possible limits as a part of the text (possibly in conclusion subsections of main chapters).

Answer: The analysis and comparison of applicability and limitations of different methods are described in included Section 4.5. In this section RNN, Kalman filter and DMD was compared in terms of accuracy and potential use cases. The advantages and disadvantages of these methods are also discussed and summarized in the table. Also, the conclusions in Sections 3.3, 3.5, 4.2, 4.3 and 4.4 were revised and updated. For Sections 3.3 and 4.3 applicability and possible limitations of CV techniques for plant phenotyping and monitoring of seeds germination were discussed. For Sections 3.5 and 4.2 applicability and possible limits for modeling and prediction of biomass using data-driven approaches were discussed. For Section 4.4 use cases and limitations for analysis of hyper-spectral data in order to find optimal wavebands for plant diseases detection were discussed.

2) Figure 1-2 contains typos in the word “engineering”

Answer: This typo was fixed and Figure 1-2 was updated in the thesis.

3) The punctuation in equations is missing. Normally, equations are considered as parts of the sentences and one expects a comma, a full stop or a continuation of the sentence after the end of an equation.

Answer: All issues with punctuation in equations were fixed.

4) Across the text there is a systematic capitalisation of first letters of the terms referring to various methods. While their acronyms are indeed made of capital letters it is not true for the full words unless they involve surnames (or copyright protected names such as Web of Science, for instance). I suggest decapitalisation of all these cases.

Answer: The thesis was carefully revised and all the cases where capitalisation is redundant were decapitalised.

5) In a few places in the text and figure captions the word “photos” is used. I would suggest a more scientifically sound “images” or “pictures”

Answer: Mainly this miswording appeared in the Section 4.3. The word “photos” was changed to “images” according your suggestions in the text and all figure captions.

6) Figure 5-1 has a term New Moscow which I would consider a spoken jargon rather than something to be used in the scientific text. I would suggest “Newly added territories in 2011”.

Answer: This terminology was changed in the Figure 5.1 and in the text in Section 5.1 according to your suggestions.

7) There are also a few typos across the manuscript to be taken care of.

Answer: The thesis was carefully revised and proofreading was done. All found typos were corrected.

Reviewer: Ulrich Schurr

1) A flaw is that the figures indicate “leaf area”, while this seems to be projected leaf area, since leaves do not shrink that massively in the diurnal cycle. Here clearly movements affect the measurement, but at the same time enrich the data set. Here I would have hoped for a more in-depth analysis of the results. The text indicates that the candidate understands the restrictions of the system (p 107 middle), but a more thorough terminology throughout the text would have been favorable.
The only issue that I consider worth correcting is the terminology with respect to leaf area and projected leaf area. As mentioned above, the candidate addresses this issue in the text, but an earlier indication to the reader and correct naming of the parameters in the figures would be appropriate.

Answer: To address this issue, more in-depth analysis was added in the Section 3.1, where the first measurements of projected leaves area are discussed. The figures and captures: 3-7, 3-24, 3-25 (a), 3-26 (a,b), 4-4 (a,b), 4-5 (a,b,c,d), 4-19, 4-20, 4-21 were fixed and updated. Also, terminology was corrected in the text in all appropriate cases, mainly in Chapters 3 and 4.

2) Chapter 5 adds another application targeted on phytotoxicity (termed “environmental parameters”), spatially explicit water quality and on testing phytotoxicology in petroleum-contaminated soils as well as the analysis of phosphogypsium on soil and plant performance. While each of these issues/ questions has a high relevance, chapter 5 gives the impression of a collection and less of targeted approach(I would have recommended to lower the number of practical examples and instead deepen some analysis of e.g. plant growth. This would have strength end the case of the thesis).

Answer: In order to present conducted research as a more targeted approach, the beginning of Chapter 5 was rewritten and Section 5.1 (Problem statement and proposed solutions) was introduced. The main aim was to implement and validate a set of data-driven approaches for improvement of the accuracy of the prediction of environmental parameters that are used in current simulators as an input data. There are two issues that should be addressed. The first issue is predicting the spatial distribution of uncontrolled environmental parameters, such as soil and water properties. The possibility of improvement of the modeling of uncontrolled parameters was shown based on the Gaussian process regression with implementation of Bayesian information criteria for automatical searching of optimal kernel structure. This approach was validated on the use case of modeling of water quality distribution. The second issue is predicting the effects on growth caused by controlled parameters. The implementation of support vector regression and artificial neural networks showed the possibility to improve the accuracy of the assessment of the effects on growth caused by inserting different amount of fertilizers and pollution.

3) The candidate then extends the analysis into early detection of pathological responses of plants using spectral analysis again the results are promising, but the applicability for the extremely wide range of plant disease would need to be addressed.

Answer: The introduction and conclusion to Section 4.4 were updated in order to show the possibility and limitations of the proposed method to be applied for a wide range of plant diseases. Overall, spectral analysis is very popular for plant diseases discrimination. The proposed technique can be used for investigation of the wide range of spectra in order to find optimal wavebands for discrimination of particular disease.

Other changes:

Fig. 3-26 was changed, to present correlations for 2 different dwarf tomato species.

Fig. 3-16 was updated, in order to represent the growth rate estimations for the period of active plant growth.

Variances of measurements that are presented in Fig. 4-4(a,b) were calculated. The error for Verhulst model fit (Fig 3-24) was calculated. Standard deviations for biomass/leaves area correlations (Fig. 3-26) were calculated.