Data-driven modeling of potato tuber size distributions

Supervisors (TU Delft): Neil Budko, Corné Verburg

Supervisors (CropConsult): Jerke Eisma

The growth and development of a potato crop is a complex dynamical process, affected by a broad range of environmental and physiological factors. The resulting size distribution is an important factor for the market value of the crop, but its development is difficult to predict due to the many interacting factors involved.

In this project, we will investigate the development of the tuber size distribution throughout the growing season as a function of environmental conditions. CropConsult, a company that provides advice to potato farmers, has collected field data containing tuber size distributions, crop measurements, weather data, soil information, and fertilization records. The goal is to use these data to identify which factors drive tuber formation and growth and to develop an interpretable dynamical model of this process.

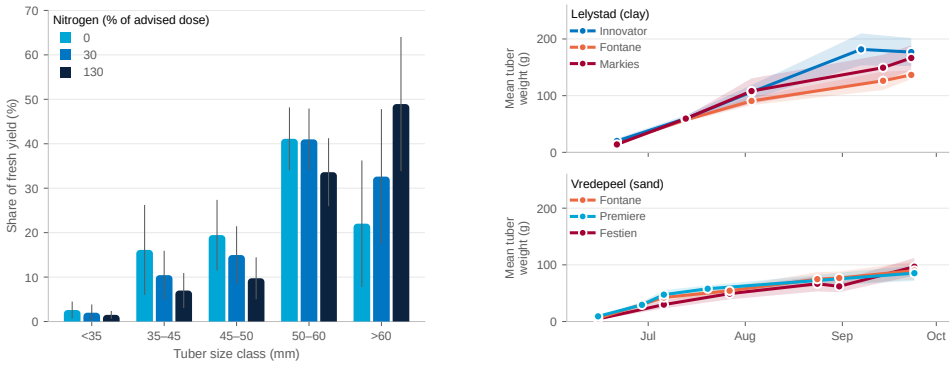


Figure 1. Share of fresh yield per tuber size class by nitrogen level (% of advised dose; left) and mean tuber weight per cultivar in 2020 (right). Whiskers and bands: ±1 SD across plots. Data: ten Den et al. (2024).

A natural mathematical framework to model crop growth is a dynamical system

$$\frac{d\mathbf{x}}{dt} = \mathbf{f}(\mathbf{x}, \mathbf{u}, t), \tag{1}$$

where $\mathbf{x} \in \mathbb{R}^n$ describes the state of the crop and $\mathbf{u} \in \mathbb{R}^m$ contains environmental inputs such as temperature, water availability, and nutrient status. To approximate $\mathbf{f}$, we use Sparse Identification of Nonlinear Dynamical Systems (SINDy), in which the dynamics are represented by a sparse combination of candidate model terms,

$$\frac{d\mathbf{x}}{dt} = \mathbf{f}(\mathbf{x}, \mathbf{u}, t) \approx \boldsymbol{\Xi}^\top \boldsymbol{\theta}(\mathbf{x}, \mathbf{u}, t), \tag{2}$$

where $\boldsymbol{\theta} \in \mathbb{R}^p$ is a vector of candidate model terms and $\boldsymbol{\Xi} \in \mathbb{R}^{p \times n}$ is a sparse coefficient matrix to be identified from data. In order to introduce variability between tubers and plants, the SINDy coefficients $\boldsymbol{\Xi}$ will be treated as random parameters, $\boldsymbol{\Xi} \sim P_{\boldsymbol{\Xi}}$.

The student will investigate how this probabilistic SINDy formulation can be used to describe the time evolution of the tuber size distribution. In particular, the student will study statistical moments such as the mean tuber size, its variance, and potentially higher-order moments. Starting from the SINDy representation, the student will derive or approximate dynamical equations for these moments and investigate suitable closure methods.

Reference for colloquium

J. Li and J. Chen, "Probability density evolution method: Background, significance and recent developments," *Probabilistic Engineering Mechanics*, 44, 111–117, 2016.

Other references

S. L. Brunton, J. L. Proctor, and J. N. Kutz, "Discovering governing equations from data by sparse identification of nonlinear dynamical systems," *Proceedings of the National Academy of Sciences*, 113(15), 3932–3937, 2016.

T. ten Den, I. van de Wiel, F. van Evert, M. van Ittersum, A. de Wit, and P. Reidsma, "Agronomic dataset on potato growth and yield in the Netherlands," *Open Data Journal for Agricultural Research*, 10, 1–13, 2024.