

least, Section 1.3 is dedicated to the introduction of hierarchical models that will further be used for the analysis of the genotypic variability of a population of plants.

1.1 General state space models

Usually the term state space models refers to linear state space models, whereas the term general state space models is used for their nonlinear equivalent. For convenience, general state space models will simply be referred to as state space models in what follows. The state space of such models can be either discrete or continuous. The SSMs considered throughout this document will be continuous-valued and discrete in time. If the equations modelling a system are continuous in time, they first need to be discretized for numerical simulation.

1.1.1 Main equations

Starting from initial state variables (initial conditions) at time step $n = 0$, the system variables are updated at each time step $n \in \llbracket 1, T \rrbracket$ where T denotes the last time step of the simulation. For plant models, this usually means that variables are updated daily, as is the case in the Log-Normal Allocation and Senescence model for *Beta vulgaris* (Section 2.2) or wheat (Section 2.3), or hourly, for instance in the GreenLab model for *Arabidopsis thaliana* (Section 2.4). At each time step n , a system of two equations summarizes the evolution of the state variables and the observations of the system respectively. In their most general form, they read:

$$\begin{cases} x_{n+1} &= f_n(x_n, u_n, \theta, \eta_n), \\ y_n &= g_n(x_n, \theta, \xi_n), \end{cases} \quad (1.1)$$

where the evolution of the system is considered between the initial time $n = 0$ and the final time $n = T \in \mathbb{N}^*$, and where at time step $n \in \llbracket 0, T \rrbracket$:

- $x_n \in \mathbb{R}^{d_x}$ represents the state variables of the model, x_0 therefore denotes the initial state of the system. Since these variables are a priori not accessible to measurement, they are also called hidden states;
- $y_n \in \mathbb{R}^{d_{y_n}}$ represents the observations on the system. It is worth noting here that the dimension of the vector of observations depends on the time step n , this will be detailed in Section 1.1.3;
- $u_n \in \mathbb{R}^{d_u}$ represents the external variables influencing the system, for example control variables: in plant science, these are typically environmental conditions in which the system evolves, such as temperature, radiation, water resources or nutrients;
- $\theta \in \mathbb{R}^{d_\theta}$ represents the functional parameters, which intervene in the functional equations and can either have a systemic meaning – they can originate from biology, for instance – or simply be parameters of empirical descriptive functions used because they constitute a convenient and sensible way of modelling a physical process;

- $\eta_n \in \mathbb{R}^{d_\eta}$ are process noises – or equivalently modelling noises – and are the values taken at each time step by a random vector representing the stochastic factors that aim to account for either possible model limitations or imperfections;
- $\xi_n \in \mathbb{R}^{d_\xi}$ are observation noises: since the observed data is most of the time measured with some uncertainty, observation noises are the values taken at each time step by the random vector defined so as to reproduce this measurement error;
- f_n is the transition function, it drives the evolution of the state variables from one time step to another;
- g_n specifies how the system is observed and what the observations are in terms of the hidden states.

The fact that both the transition and observation functions are allowed to depend on the time index n is referred to as non-homogeneous transitions. This is often the case in plant growth models since the plant has different evolution stages (which mostly depends on the thermal time) where its behaviour can be drastically different.

1.1.2 Hidden Markov models

In their stochastic formulation with random vectors defining the process and observation noises, SSMs are equivalent to hidden Markov models (HMMs) [Rabiner, 1989] where x_n represents the hidden states, y_n the observations and where:

$$\begin{cases} x_0 & \sim p(x_0) & \text{is the initial distribution,} \\ x_{n+1} & \sim p(x_{n+1}|\theta, x_n) & \text{is the transition distribution,} \\ y_n & \sim p(y_n|\theta, x_n) & \text{is the observation distribution.} \end{cases} \quad (1.2)$$

The transition and observation distributions represented above by conditional probability density functions can be rewritten using the process and observation noises as will be detailed in Section 1.2. It has to be noted that each one of these distributions can be taken as a Dirac distribution. An important case is that of a model without process noise, in which case Equation 5.2 reduces to:

$$\begin{cases} x_{n+1} & = f_n(x_n, u_n, \theta), \\ y_n & = g_n(x_n, \theta, \xi_n). \end{cases} \quad (1.3)$$

In this case, the transition distribution is equivalent to a Dirac distribution so that $p(x_{n+1}|\theta, x_n)dx_{n+1}$ is replaced by $\delta_{f_n(x_n, u_n, \theta)}(dx_{n+1})$ and for given parameters θ , initial state x_0 and external variables $(u_n)_{n \in \llbracket 1, T \rrbracket}$, all state variables at all time steps $(x_n)_{n \in \llbracket 1, T \rrbracket}$ are deterministically defined. A less common case would be that of a model without observation noise, where $p(y_n|\theta, x_n)dy_n$ would be replaced by $\delta_{g_n(x_n, \theta)}(dy_n)$, which would be such that every measurement contains perfect information on the system. This scenario, however, finds very few practical applications as most models deal with the measurement of continuous valued variables and thus necessarily involves some uncertainty.

1.1.3 Structure of the observations

At a given time step n , the observed variables can be of different nature: integers (the number of phytomers of a plant), real numbers (the biomass of a leaf compartment), vectors (the areas of the different individual leaves) or even matrices in some cases. Most of the time and for practical reasons in real world applications, observations do not come up at every time step and the data available at a given time might not be the same at another. Let us consider the example of a plant model for, say, sugar beet: the biomass of the leaves might be available on days 10, 20 and 35 whereas the biomass of the roots might be available on days 12, 20, 33. For an organ-scale plant model where observations on the different organs are obtained via image analysis, as is the case of leaf areas, the latter might not be available on the same days because an algorithm deemed its classification confidence to be insufficient. The size and content of the observations might therefore not be the same through time, which poses no problem whatsoever as long as one knows what variables are observed at what time. The merged experimental timeline, representing time steps at which any experimental data is available, will be denoted by:

$$\mathcal{O} = (t_k)_{k \in \llbracket 1:O \rrbracket} \in \mathbb{N}^O \text{ with } 1 \leq t_k < t_{k+1} \leq T \text{ for } k \in \llbracket 1, O \rrbracket, \quad (1.4)$$

where $O \geq 1$ is the total number of experimental time steps. More formally, the observations at a given time step n can be seen as a dictionary where keys would be the different observed variables through the experiment and the values would be the actual observations of the corresponding variables. The operation of converting a dictionary of observations into a vector and its inverse will be presented in more detail in Section 5.4 when discussing the practical implementation of the storage of observations using the computing platform. If ℓ_n denotes the total number of variables observed at time step n , $v_n = (v_n^\ell)_{\ell \in \llbracket 1, \ell_n \rrbracket}$ said variables and y_n^ℓ denotes the observation relative to variable v_n^ℓ , then the vector of observations at time n can be defined as the concatenation:

$$y_n = (y_n^\ell)_{\ell \in \llbracket 1, \ell_n \rrbracket} \in \mathbb{R}^{d_{y_n}}. \quad (1.5)$$

Equivalently, all the observation vectors at each time in the experimental timeline can be concatenated and the following notations are introduced:

$$\begin{cases} x_{1:T} &= (x_n)_{n \in \llbracket 1:T \rrbracket}, \\ y_{1 \rightarrow T} &= (y_{t_k})_{k \in \llbracket 1:O \rrbracket}. \end{cases} \quad (1.6)$$

Some algorithms (least squares algorithms for example) require to deal with vector observations and it is therefore important to be able to work with observations of such a nature. In fact, $y_{1 \rightarrow T}$ could also be seen as a matrix of observations where each row would correspond to a time step and each column to a type of observation, and elements of this matrix could be missing (since all variables are not observed at all times). All the information about what variables are observed at what time is actually stored in the sequence of observation functions $(g_n)_{n \in \llbracket 1:T \rrbracket}$. More details on how this is implemented in the computing platform can be found in Sections 5.2.3 and 5.3 with examples. For now, one can assume that at each time step n , y_n is a vector of observations, that are not necessarily the same at different times, and that one knows what these vector observations contain and how to exploit them.

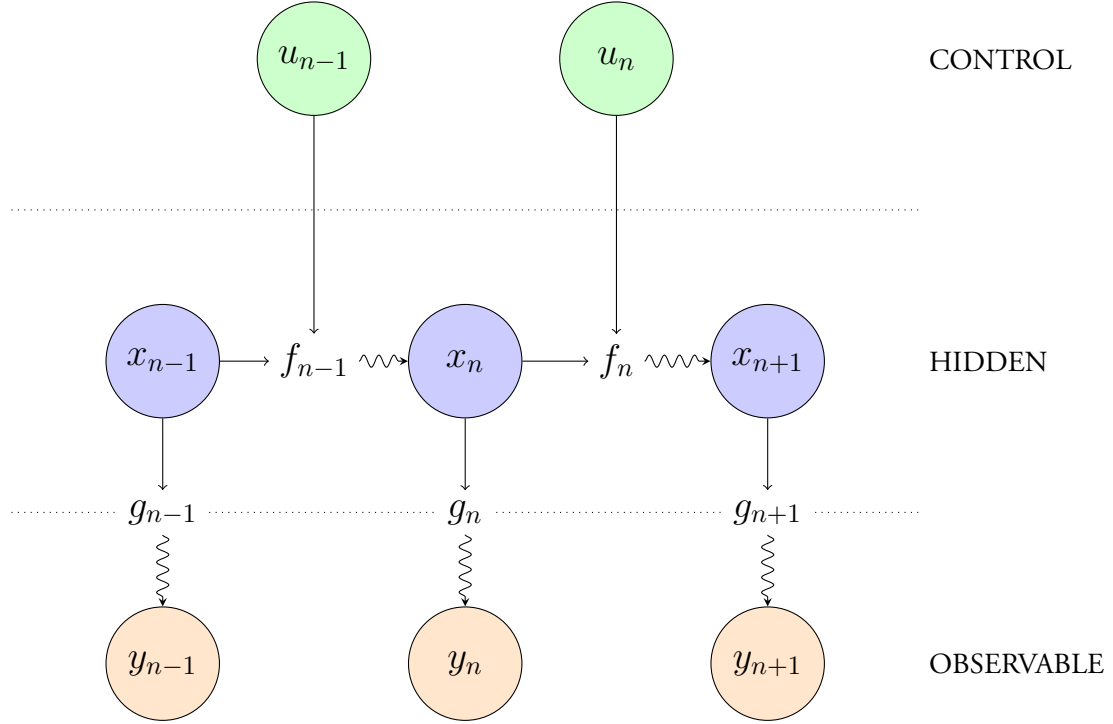


Figure 1.1: Representation of a general state space model. Wavy curves in the hidden layer (respectively observable layer) represent the randomness introduced by the process noise (respectively observation noise). The external (or control) variables u_n are known at every time step, the hidden states x_n are unknown in real experiments but are accessible when the corresponding model is simulated, and the observations y_n are both known in real experiments and can also be simulated provided that an observation model error has been defined.

Process and observation noises are of stochastic nature, and the underlying parameters constant throughout a model simulation are typically the mean or standard deviation of the statistical distribution from which they are sampled. We distinguish the random variables η and ξ , from their realizations at a given time η_n and ξ_n such that:

$$\begin{cases} \eta & : & \Omega^\eta & \rightarrow & \mathbb{R}^{d_\eta} \\ \omega^\eta & \rightarrow & \eta_n \equiv \eta(\omega^\eta) \end{cases} \quad (1.7)$$

and:

$$\begin{cases} \xi & : & \Omega^\xi & \rightarrow & \mathbb{R}^{d_\xi} \\ \omega^\xi & \rightarrow & \xi_n \equiv \xi(\omega^\xi) \end{cases} \quad (1.8)$$

where Ω^η and Ω^ξ are appropriate sample spaces. A simulation of the model can therefore be summarized as:

$$y_{1 \rightarrow T} = M(x_0, u, \theta, \eta, \xi). \quad (1.9)$$

where the model M contains the information on the sequences of transition and observation functions $(f_n)_{n \in \llbracket 1, T \rrbracket}$ and $(g_n)_{n \in \llbracket 1, T \rrbracket}$. Sometimes, one might abbreviate the output of the model as $y \doteq y_{1 \rightarrow T}$. It is worth noting that the use of such a stochastic formulation for plants is not very common and dates back to less than 20 years [Makowski et al., 2004], [Chen and Cournède, 2012], [Trevezas and Cournède, 2013]. A graphical representation of a general SSM is shown on Figure 1.1.

1.1.4 Observation model

Considering a plant model, the biomass produced by the whole plant on a given day is not directly measurable, it is thus considered to be a hidden state, whereas the biomass of green leaves on the same day is a measurable data, making it an observation. As previously said, a measure on a system is very often inexact, and this is almost always the case when dealing with continuous-valued models. For instance, measuring biomass can be done using either destructive or non-destructive methods: in the first case, biomass is removed from the plant and weighed while the second case is based on digital image analysis. The cutting point, the weighing process, or imperfections of algorithms are as many factors implying that some measurement error is made. The true value of the biomass is therefore never exactly measured, and a distinction must be made between the hidden state of the biomass, which remains unknown, and its corresponding observation. A given measurable variable will frequently be defined both as a hidden state and as an observation. In the case of a biomass denoted by q , this would translate into:

$$q_n \in x_n \text{ and } \tilde{q}_n \in y_n. \quad (1.10)$$

How these two values are related constitute a model for the measurement error. A standard approach is to consider that, on average, the hidden state is measured with some white noise following a normal distribution. If the noise is proportional to the value of the hidden state – for instance if the greater the biomass, the greater the measurement error – one might want to consider a multiplicative noise:

$$\tilde{q}_n = q_n (1 + \xi_n), \text{ with } \xi_n \sim \mathcal{N}(0, (\sigma^q)^2) \text{ and } \sigma^q > 0, \quad (1.11)$$

so that:

$$\tilde{q}_n \sim \mathcal{N}\left(q_n, (\sigma^q q_n)^2\right), \quad (1.12)$$

whereas if the noise does not depend on the value of the hidden state, one might want to consider an additive noise:

$$\tilde{q}_n = q_n + \xi_n, \text{ with } \xi_n \sim \mathcal{N}(0, (\sigma^q)^2) \text{ and } \sigma^q > 0, \quad (1.13)$$

so that:

$$\tilde{q}_n \sim \mathcal{N}\left(q_n, (\sigma^q)^2\right). \quad (1.14)$$

Obviously, there can be situations where measured values are always underestimated or overestimated, in which case these two measurement error models might not be relevant anymore. In the rest of this thesis, the values that will be measured will be either biomasses or leaf areas, and it is therefore assumed that multiplicative normal noises are the most adapted to such situations. However, more observation models have been considered as will be emphasized in Chapter 5 when discussing the computing platform.

1.1.5 Extensions

The popularity of such models has generated many extensions. A common one concerns k -th order Markov process, where $k > 1$: in this case the hidden state x_{n+1} does not depend only on x_n but on $(x_{n-j+1})_{j \in \llbracket 1, k \rrbracket}$.

This can happen in plant growth models – such as in the STICS model [Brisson et al., 1998] for instance – even though from a mathematical point of view it is always possible to redefine the system state as $x'_n = (x_{n-j+1})_{j \in \llbracket 1, k \rrbracket}$ to define a standard SSM again.

In Markov-switching models (also called Markov jump systems), at time step n the observation y_n depends not only on the hidden state x_n but also on the previous observation y_{n-1} (and possibly on even older observations) [Cappé et al., 2005]. The sequence of observations $(y_n)_{n \in \llbracket 1, T \rrbracket}$ can therefore be seen, conditional on the sequence of hidden states $(x_n)_{n \in \llbracket 1, T \rrbracket}$, as a non-homogeneous Markov chain. Although this kind of model has a lot in common with standard HMMs, their statistical analysis is much more complicated because the observed sequence $(y_n)_{n \in \llbracket 1, T \rrbracket}$ is not directly related to that of the unobservable one $(x_n)_{n \in \llbracket 1, T \rrbracket}$. Although Markov-switching models are not considered within this thesis, two reasons behind their potential uses must be mentioned. First, when measurements are performed on a plant or in a field, the observer might be influenced by the previous obtained results: if the biomass of an organ is surprisingly much lower than that of a previous time, one might be tempted to correct the current measurement upwards. Second, the image analysis algorithm used to estimate the individual leaf areas (see Chapter 7) does so by taking into account the whole history of a given leaf. For the image of a given day, the decision to classify a given segment (i.e. a set of connected pixels that was considered to represent a leaf occurrence) – whose area contains some observation noise intrinsic to the segmentation algorithm – as belonging to a particular leaf of the plant depends on the whole history of the said leaf, hence on previous observations. Nevertheless, this effect is considered to be of minimal importance in the present case: when it is uncertain whether a segment belongs to a leaf, it is considered as not being observed. Classification-related errors (and time-induced) will therefore be minimal in the pool of actually observed data collected.

1.2 Generic probability distributions

As will be detailed in Chapter 3, parameter and state estimation algorithms require the use of the transition and observation probability density functions (pdfs). The algorithms should be designed so as to be easily used with different models. It therefore requires the calculation of the transition pdf $p(x_{n+1}|\theta, x_n)$ and the observation pdf $p(y_n|\theta, x_n)$. What is more, the pdf of all the observations conditional to the parameters and the hidden states $p(y_{1:T}|\theta, x_{1:T})$ can then easily be deduced from the observation pdf. For this aim, a generic expression of the latter is derived and, as explained in Chapter 5, this will allow to automatically compute their values provided that models are written using a predefined template.

1.2.1 Transition probability density function

In the models considered, it is always possible to arrange the state variables by their order of computation at a given time step. The hidden state is therefore decomposed into $(x_n^j)_{j \in \llbracket 1, d_x \rrbracket}$ where for all $j \in \llbracket 1, d_x - 1 \rrbracket$, x_n^j is computed before x_n^{j+1} . In particular, a variable x_{n+1}^j can depend in practice on all variables computed before, $(x_{n+1}^k)_{k \in \llbracket 1, j-1 \rrbracket}$ – which have been expressed as functions of x_n themselves – without breaking the

dependence on only x_n from a mathematical point of view. The transition pdf can therefore be expressed in a hierarchical fashion:

$$p(x_{n+1}|\theta, x_n) = \prod_{j \in \llbracket 1, d_x \rrbracket} p(x_{n+1}^j|\theta, x_n, x_{n+1}^{1:j-1}). \quad (1.15)$$

In the absence of process noise, the dynamics of the system is entirely deterministic, as a consequence $p(x_{n+1}^j|\theta, x_n, x_{n+1}^{1:j-1})dx_{n+1}^j = \delta_{m_{n+1}^j(x_n, x_{n+1}^{1:j-1}, \theta)}(dx_{n+1}^j)$ where $m_{n+1}^j(x_n, x_{n+1}^{1:j-1}, \theta)$ prescribes how x_{n+1}^j is computed within the model. In a model without process noise, the transition pdf therefore becomes such that:

$$p(x_{n+1}|\theta, x_n)dx_{n+1} = \prod_{j \in \llbracket 1, d_x \rrbracket} \delta_{m_{n+1}^j(x_n, x_{n+1}^{1:j-1}, \theta)}(dx_{n+1}^j). \quad (1.16)$$

In particular, one can choose to introduce as many intermediary variables in the hidden state x_n without fundamentally changing the model formulation in terms of transition distribution. In view of this remark, when some process noise is involved, the only non-Dirac terms in Equation 1.15 are those affected by the process noise and the corresponding random vector can also be arranged by order of use in the model $\eta_n = (\eta_n^j)_{j \in \llbracket 1, d_\eta \rrbracket}$. For now, we assume that all the process noises of the model are unidimensional.

Let $m_\eta : \llbracket 1, d_\eta \rrbracket \rightarrow \llbracket 1, d_x \rrbracket$ be an application such that $m_\eta(\llbracket 1, d_\eta \rrbracket)$ represents the sets of indices of the state variables on which are set the process noises, i.e. there exists some function ϕ_j such that:

$$x_n^{m_\eta(j)+1} = \phi_j(x_n^{m_\eta(j)}, \eta_n^j), \text{ for } j \in \llbracket 1, d_\eta \rrbracket. \quad (1.17)$$

For a particular process noise of index j and in the case of an additive normal noise, this would translate into:

$$x_n^{m_\eta(j)+1} = x_n^{m_\eta(j)} + \eta_n^j \text{ with } \eta_n^j \sim \mathcal{N}(0, (\sigma^j)^2). \quad (1.18)$$

It is possible to express the transition pdf by choosing only the d_η state variables on which are set the process noises:

$$p(x_{n+1}|\theta, x_n)dx_{n+1} = \prod_{j=1}^{d_\eta} p(x_{n+1}^{m_\eta(j)+1}|\theta, x_{n+1}^{m_\eta(j)})dx_{n+1}^{m_\eta(j)+1} \times \prod_{j \notin m_\eta(\llbracket 1, d_\eta \rrbracket)} \delta_{m_{n+1}^j(x_n, x_{n+1}^{1:j-1}, \theta)}(dx_{n+1}^j)$$

where we recall that $m_{n+1}^j(x_n, x_{n+1}^{1:j-1}, \theta)$ is the value computed for the state variable x_{n+1}^j within the model considered. In what follows, since the variables that are deterministic functions of the stochastic variables are computed directly by closure relationships in the simulation program, the Dirac distributions take values 1, hence will be omitted in the following. In particular, we can restrain ourselves to the noised variables and replace the transition pdf by:

$$p(x_{n+1}^{m_\eta(j)+1}, \dots, x_{n+1}^{m_\eta(d_\eta)+1}|\theta, x_n) = p(\eta_n|\theta) = \prod_{j=1}^{d_\eta} p(x_{n+1}^{m_\eta(j)+1}|\theta, x_{n+1}^{m_\eta(j)}). \quad (1.19)$$

For the sake of simplicity and with a slight abuse of notation, in what follows we will denote $p(x_{n+1}|\theta, x_n) = p(x_{n+1}^{m_\eta(j)+1}, \dots, x_{n+1}^{m_\eta(d_\eta)+1}|\theta, x_n)$, implicitly assuming Dirac distributions for the variables computed in a deterministic fashion by model closure. This formulation generalizes in fact very well to multidimensional noises. The decomposition of the state variable $x_n = (x_n^j)_{j \in \llbracket 1, d'_x \rrbracket}$ can be performed in such a way that x_n^j can be a multidimensional quantity such as a vector or a matrix and so that $d'_x \leq d_x$.

Similarly, $\eta_n = (\eta_n^j)_{j \in \llbracket 1, d'_\eta \rrbracket}$ can contain multidimensional quantities. For instance, the random vector of process noises in the Kalman filter [Kalman, 1960] is drawn from a multivariate normal distribution with non-zero off-diagonal components (i.e. this cannot be simplified to the use of unidimensional process noises), which for some j would translate into:

$$x_n^{m_\eta(j)+1} = x_n^{m_\eta(j)} + \eta_n^j \text{ with } \eta_n^j \sim \mathcal{N}(0, \Sigma) \quad (1.20)$$

and where Σ can potentially be a full matrix.

1.2.2 Observation probability density function

Equivalently, one can decompose the observation noises as $\xi_n = (\xi_n^j)_{j \in \llbracket 1, d'_\xi \rrbracket}$ and define an application $m_\xi : \llbracket 1, d'_\xi \rrbracket \rightarrow \llbracket 1, d'_x \rrbracket$ such that $m_\xi(\llbracket 1, d'_\xi \rrbracket)$ represents the set of indices of the state variables on which are set the observation noises, i.e. there exists some function ψ_j such that:

$$y_n^j = \psi_j(x_n^{m_\xi(j)}, \xi_n^j), \text{ for } j \in \llbracket 1, d'_\xi \rrbracket. \quad (1.21)$$

Again, for a particular observation noise of index j and in the case of a unidimensional multiplicative noise, this would translate into:

$$y_n^j = x_n^{m_\xi(j)} (1 + \xi_n^j) \text{ with } \xi_n^j \sim \mathcal{N}(0, (\sigma^j)^2). \quad (1.22)$$

It can also happen that observations on the system require multidimensional noises. This is notably the case in [Baey et al., 2016] where the biomasses of the different organs are observed with correlation, this would mean that:

$$y_n^j = x_n^{m_\xi(j)} \cdot (1 + \xi_n^j) \text{ with } \xi_n^j \sim \mathcal{N}(0, \Sigma). \quad (1.23)$$

where it is understood that in the case of multidimensional noises, operations on vectors such as \cdot or $/$ are performed element-wise, and with Σ having some of its off-diagonal components non-zero. The observation pdf can be expressed in the same way as for the transition pdf:

$$p(y_n | \theta, x_n) = p(\xi_n | \theta) = \prod_{j=1}^{d_{\xi'}} p(y_n^j | x_n^{m_\xi(j)}) \quad (1.24)$$

where the product runs on all indices j for which experimental data y_n^j is available and, again, possibly contains multidimensional noises. The generic expression of the pdf of the observations conditional to the parameters and the hidden states $p(y_{1 \rightarrow T} | \theta, x_{1:T})$ naturally follows from that of the observation pdf since:

$$p(y_{1 \rightarrow T} | \theta, x_{1:T}) = \prod_{k=1}^O p(y_{t_k} | \theta, x_{t_k}) = \prod_{k=1}^O \prod_{j=1}^{d_{\xi'}} p(y_{t_k}^j | \theta, x_{t_k}^{m_\xi(j)}). \quad (1.25)$$

Equation 1.19 makes it possible to compute the transition pdf as long as are specified the nature of the noises (are they additive or multiplicative, sampled from a normal, a log-normal or a uniform distribution?) and lists of labels corresponding to the parameters necessary to compute the values of these pdfs. Likewise, Equation 1.24 makes it possible to compute the observation pdf. The complete mechanism for the practical computation of such values of the transition and observation pdfs are described in detail in Section 5.5.

Examples of theoretical transition and observation pdfs are given in Chapter 2 in the case of two plant growth models for *Beta vulgaris* and *Arabidopsis thaliana*.

1.3 Population models

In this section, we first describe the reasons behind the use of the population approach and emphasize its importance in the context of genotypic variability. We then move on to introduce the two types of variability in the context of plants and introduce hierarchical models, statistical models that are well-suited for dealing with this dual variability for a population of plants, and finally show how the SSMs considered in Section 1.1 fit in the mathematical description of population models before describing the complete two-stage hierarchical models that will further be used for parameter inference in a population context.

1.3.1 Motivation

Most biological and physical phenomena observed within a set of different individuals exhibit variability: they might manifest similar global behaviour and dynamics but with some variations. This is of practical use in many fields such as biology, agronomy, econometrics, environmental and human sciences. For instance, in pharmacometrics, one needs to develop models where different patients would react differently to the same disease and the same drug. [Lavielle \[2014\]](#) gives the example of the effect of genetically modified corn on the health of rats.

As far as plants are concerned, there exists a strong genetic variability even within individuals of the same species, which notably allows for better resistance to diseases or bugs and provides stronger adaptation to a wide set of environmental conditions. [Brouwer et al. \[1993\]](#) showed that soil and crop growth micro-variability in the semi-arid tropics of West Africa contributes to increased yield in case of droughts since parts of the field, more resistant to water stress, could compensate for other parts performing poorly, meaning a satisfactory level of assured production. The genetic variability of switchgrass was studied by [Hopkins et al. \[1995\]](#) in order to develop improved populations, which highlighted the importance of genotype-environment interactions for traits such as forage yield at heading, vegetative in vitro dry matter digestibility and heading date. For maize, [Maiti et al. \[1996\]](#) showed that both genotypic variability and soil content were of highly significance for the resistance to drought and salinity at the seedling stage. This study also allowed to speculate about the effect of higher root growth under saline stress in some of the genotypes as a mechanism of resistance in maintaining osmo-regulation. [Isfan \[1993\]](#) suggested that the index of physiological efficiency of absorbed nitrogen may be used in order to identify the likely high yielding oat genotypes and those capable of exploiting nitrogen input most efficiently.

All these examples indicate both the importance of such a genetic variability and the necessity to integrate it within plant growth models. Similarly, in populations of plants in fields or forests, interindividual variability can result from differences in the local environmental conditions of individual plants. The first plant growth

models integrating this variability tried to simulate the growth of each individual with a competition index between plants [Fournier and Andrieu, 1999], [Cournède et al., 2008].

Most of the time in plant applications, repeated measurements on different individuals of a population are available, and a population approach is particularly suited to characterize and explain this kind of data. The mathematical model used therefore needs to incorporate a growth model that depicts the dynamics of the different state variables of the plant considered – biomasses mainly – and a statistical model that explains the variations of this typical dynamics between the different individuals. There are in fact two sources of variability to account for:

- the intraindividual variability, which refers to how the state of a single individual might vary, because of random processes and measurement errors. This kind of variability was introduced in Section 1.1 under the form of process and observation noises;
- the interindividual variability, which arises because of differences between the genotypes or environments of different individuals. This corresponds to the variability of the different individuals' curves around a mean population curve.

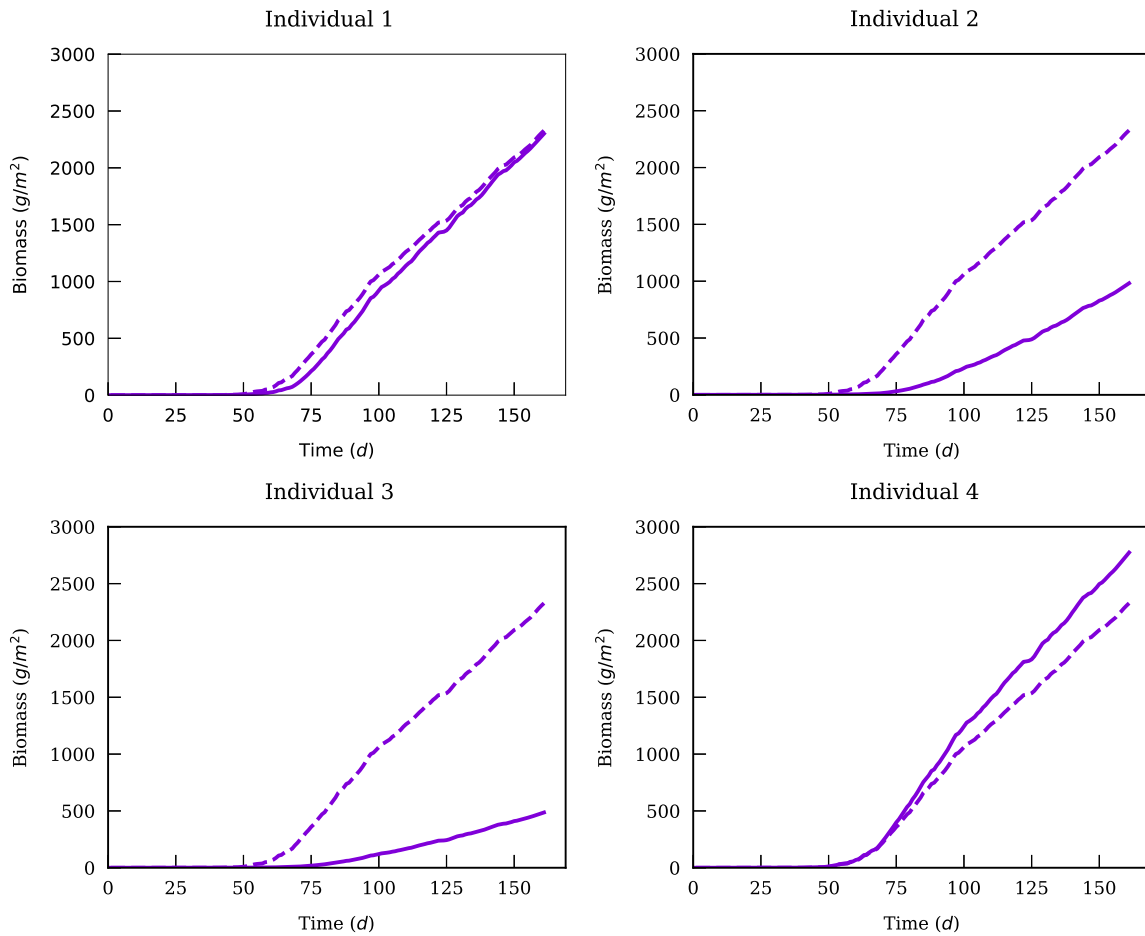


Figure 1.2: Yield curves of four different individuals. The individual curves are displayed with solid lines and the mean curve (similar for all graphs) is displayed with dotted line.

1.3.2 Hierarchical models

Hierarchical models provide a very suitable way of modelling this dual variability. This type of statistical models have long been used in pharmacokinetics, epidemiology or ecology, although their use in dynamic plant growth models is rather recent [Baey et al., 2013]. Broadly speaking, they are statistical models of parameters that vary at several levels. First introduced in the context of linear regression, multilevel models then consisted in doing a regression in which the parameters were given a probability model, and this second-level model had parameters of its own, called hyperparameters, and all these parameters were inferred from the same data.

Many names can be found for such models: multilevel models may be the more common, nested data models, random effects models (mixed effects models when some parameters are random and some are fixed in the population) are also widely used. The hierarchical term seemed the most appropriate to our case since a hierarchy is clearly established between the different stages of plant growth modelling where individual parameters are first derived from a given probability distribution describing the population, and these individual parameters then drive the growth of a given plant via a noisy nonlinear SSM.

The two most simple and direct methods of parameter estimation in a population are known as complete pooling and no pooling [Davidian and Giltinan, 1993]:

- in complete pooling, differences between individuals are ignored, all are treated equally and the data coming from different individuals is used for the estimation of parameters supposed to represent all individuals. Disregarding the diversity of the sources for the data represents a significant oversimplification, not only will it fail to provide accurate predictions but it also misses variations in the population, which is the main objective of many studies;
- in no-pooling, data coming from different sources are analyzed separately. This procedure ignores the information related to the diversity of the individuals in the population and can lead to unsatisfactorily variable inferences for the model parameters, in particular when little data is available. It must be noted, though, that individual estimates can still be of use for the initialization of prior of hyperparameters in a Bayesian estimation procedure as will be described in Section 8.1.2.

Hierarchical modelling manages to combine the information provided by different individuals and overcome the limitations introduced by these two simplistic methods.

The first stage of the model corresponds to the intraindividual variability and aims to explain how an individual evolves given a set of parameters supposed to represent it. The dynamics of each individual i within the population is then described by the same parametric model with a different set of parameters θ_i . Generalized linear models are often used for convenient and straightforward inference, however, the highly nonlinear nature of plant growth models suggested to proceed otherwise, and this first stage will be represented by the dynamics of general SSMs described in Section 1.1. The typical dynamics of an individual is thus well described with a parametric model such as those presented in Chapter 2 and a mean population dynamics y is assumed to be obtained with a set of parameters θ . The second stage of the model corresponds to the

interindividual variability and depicts the common probability distribution of the different sets of parameters that control the time dynamics of the individuals. The interindividual variability of the curves around the typical population curve can therefore be explained by the parameter variability around the mean population parameters as can be observed on Figure 1.2 for the yield curves of 4 different individuals. The parameters of the model can be considered either fixed or random. The random parameters are assumed to follow a statistical distribution parameterized by the typical population parameters and individual parameters are therefore random variables following the same population distribution. In a Bayesian framework, a third stage is finally added for the prior distribution of the population parameters.

1.3.3 State space models for populations

For the sake of consistency with literature, the mathematical notations used in the population approach will slightly differ from the single individual case. A population is made up of N different individuals indexed by $i \in \llbracket 1, N \rrbracket$. If $\theta_i \in \mathbb{R}^{d_\theta}$ denotes the set of parameters for individual i , the state space Equations 1.1 become in their most general form:

$$\begin{cases} x_{i,n+1} &= f_n(x_{i,n}, u_{i,n}, \theta_i, \eta_{i,n}), \\ y_{i,n} &= g_n(x_{i,n}, \theta_i, \xi_{i,n}). \end{cases} \quad (1.26)$$

Hopefully, some simplifications can be made. In the population approach, the nature of variability is two-fold and can be assumed, for the sake of simplicity, to encompass randomness arising from both process and observation noises. The process noise $\eta_{n,i}$ will therefore be ignored. As all applications of this thesis will be done within controlled environments, the control variables $u_{i,n}$ will be the same for all individuals and can be omitted. The transition part of Equation 1.26 can thus be simplified to:

$$x_{i,n+1} = f_n(x_{i,n}, \theta_i) \quad (1.27)$$

which allows, by induction, to rewrite more simply for all n :

$$x_{i,n+1} = h_n(x_{i,0}, \theta_i), \quad (1.28)$$

or in an even simpler form by incorporating the initial state into the h_n function – or the parameter vector if it is unknown:

$$x_{i,n+1} = h_{i,n}(\theta_i). \quad (1.29)$$

1.3.4 Intraindividual variability

As previously mentioned, when dealing with population models in the rest of this thesis, it will always be assumed that all observed variables follow a multiplicative normal observation model, which means that the measurement error associated to a given state is proportional to the latter. Note however that this assumption is made without loss of generality from a methodological point of view. For each individual, there are n_i measurements indexed by $j \in \llbracket 1, n_i \rrbracket$ and y_{ij} therefore denotes the j -th measurement on individual i , although index j does not designate time per se. The observations for the different individuals need not

be at the same times. In the context of population models, the vector of all observations from $n = 1$ to $n = T$ for a given individual i is denoted by $y_i = (y_{ij})_{j \in \llbracket 1, n_i \rrbracket} \in \mathbb{R}^{n_i}$. Although the notations look similar, there is usually little confusion possible between the observations y_n at time n for a single individual and the whole vector of observations y_i from $n = 1$ to $n = T$ for a given individual i . Given the complexity of the observations in the case of the GreenLab model for *Arabidopsis thaliana* within a population approach, details about the conversion between the observations at each time and the concatenated vector of all observations for the whole simulation are provided in Section 2.4. Assuming that the hidden state corresponding to the j -th observation of the i -th individual is simulated and denoted by $h_{ij}(\theta_i)$, then the observation model considered implies that:

$$y_{ij} = h_{ij}(\theta_i) \times (1 + \xi_{ij}) \text{ with } \xi_{ij} \sim \mathcal{N}(0, \sigma^2). \quad (1.30)$$

More elaborate observation models specifying heterogeneous observation-related variances, such as the one proposed in [Duval et al., 2009] could also be considered. It is worth noting that the standard deviation associated to the observation noise depends neither on the individual nor on the observations within a given individual. With the standard notations of hierarchical models, this becomes:

$$y_{ij} \sim \mathcal{N}(h_{ij}(\theta_i), \sigma^2 h_{ij}(\theta_i)^2) \quad (1.31)$$

for $i \in \llbracket 1, N \rrbracket$ and $j \in \llbracket 1, n_i \rrbracket$. Another way of rewriting this model with the vector of all observations of the i -th individual y_i reads:

$$y_i = \begin{pmatrix} y_{i1} \\ \vdots \\ y_{in_i} \end{pmatrix} = \begin{pmatrix} h_{i1}(\theta_i) + \xi_{i1} h_{i1}(\theta_i) \\ \vdots \\ h_{in_i}(\theta_i) + \xi_{in_i} h_{in_i}(\theta_i) \end{pmatrix} \sim \mathcal{N}(h_i(\theta_i), \tau^{-1} \Omega_i) \quad (1.32)$$

where $h_i(\theta_i) \in \mathbb{R}^{n_i}$ is the vector of the hidden states corresponding to the experimental data for the i -th individual (given by the model), and $\tau = \sigma^{-2} \in \mathbb{R}^{*+}$ is called the precision. The reason for using the precision instead of the standard deviation will become apparent when prior distributions for the estimation of parameters in a Bayesian framework are discussed in Chapter 4. For the multiplicative observation model considered, it is straightforward that Ω_i reduces to:

$$\Omega_i = \text{diag}\{h_i(\theta_i)^2\}. \quad (1.33)$$

More complicated covariance matrices could be used in practice, but this is not relevant to the case considered thereafter: each measurement is considered independant of the others, whence the diagonal matrix.

1.3.5 Interindividual variability

As the first stage of the model describes the intraindividual variability, the second stage of the model deals with the interindividual variability and prescribes how the individual parameters θ_i are distributed within the population. One of the simplest yet most sensible way to do so is to consider that the individual parameters $\theta_i \in \mathbb{R}^{d_\theta}$ follow a normal distribution:

$$\theta_i \sim \mathcal{N}(\eta, \Sigma) \quad (1.34)$$