HAL
open science

Study of the ranges of values of a Biomimetic Statistical Learning Tool parameters

Hélène Flourent

To cite this version:

Hélène Flourent. Study of the ranges of values of a Biomimetic Statistical Learning Tool parameters. 2019. hal-02067374

HAL Id: hal-02067374

https://hal.science/hal-02067374v1

Preprint submitted on 14 Mar 2019

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Study of the ranges of values of a Biomimetic Statistical Learning Tool parameters.

Hélène Flourent[a]

[a] *Université Bretagne Sud, Laboratoire de Mathématiques de Bretagne Atlantique,UMR CNRS* 6205, *Campus de Tohannic, Vannes, France*

Abstract

We built a *Biomimetic Statistical Learning Tool* based on a *PDE* system, embarking the mathematical expression of biological determinants, in order to perform Data Fitting and Data Assimilation (See Flourent *et al.* (Submitted)).

Before starting the fitting of the parameters of a model, we have to specify for each parameter a lower and an upper values between which the optimization algorithm will search the value minimizing the error associated to the model.

The goal of this working paper is to detail the study of the ranges of values of the parameters of the built *Biomimetic Statistical Learning Tool*.

Keywords: Range of Values, Statistical Learning Tool, PDE, Forecasting, Data Assimilation, Model-Data Coupling, Mathematical Modeling, Biological phenomena

1. Introduction

In Flourent *et al.* (Submitted) a *Biomimetic Statistical Learning Tool* based on a *PDE* system, embarking the mathematical expression of biological determinants, was built. The objectives of this tool is to perform Data Fitting and Data Assimilation.

During Data Fitting, an optimization algorithm search the model parameter values minimizing the error of the model on a *Training Database*, between an upper and a lower bound. Therefore we have to delimit a priori the ranges of values of the parameters to fit. The objective of this working paper is to establish a methodology permitting to determine the range of values of each parameter.

2. The studied *Mathematical Model*

2.1. Description of the Mathematical Model

For self-containedness of this working paper, we recall in the following paragraph almost word-for-word the explanation of the *Mathematical Model* given in Flourent *et al.* (Submitted).

The *Mathematical Model* developed in Flourent *et al.* (Submitted), mathematically traduces the evolution and the action of a global information circulating in an *Avatar* of a *Real Animal*. The authors made the assumption that, when an active or a nutrient

enters the body of a living organism, it circulates in the body through a network of vessels containing a fluid. This element integrates this fluid and evolves via convection and diffusion phenomena. In this network of vessels the element may be in competition with others which may delay its forward progression. Then the circulating element may be caught, accumulated and then used to induce change in some biological variables. We can mathematically traduce all those processes through a PDE system which is illustrated by Figure 1.

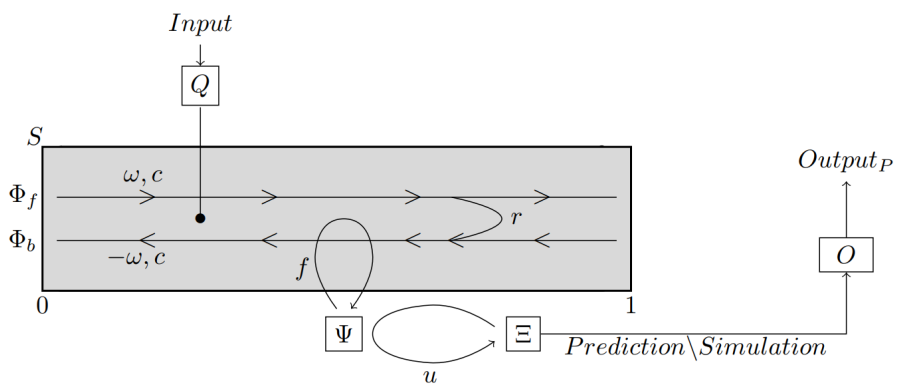

Figure 1: Schematization of the *Mathematical Model* (Flourent *et al.* (Submitted))

The developed system of Partial Differential Equations contains several parameters that have to be learnt. There are $\omega_{\mathtt{d}}$ and $c_{\mathtt{d}}$, the convection and the diffusion speeds, $f_{\mathtt{d}}$ and $u_{\mathtt{d}}$, the fixation and the usage coefficient and $r_{\mathtt{d}}$, the delay parameter. See Flourent *et al.* (Submitted) for further detail.

Therefore, the ranges of values of those five parameters have to be determined. We already know that all the parameters are positive, hence the lower bound of the different ranges of values is zero.

2.2. Discretization of the Mathematical Model

To discretize the *Mathematical Model* the classical Finite Difference method was used to obtain semi-discrete in space equations. Then R-function *Ode.1D* developped by Soetaert *et al.* (2010) was used to manage the discretization in time of the semi-discrete equations. This R-function calls upon the fourth order Runge Kutta method with a given time step (See Enright (1989)).

3. Determination of the range of values of the convection and the diffusion speeds

The convection speed $\omega_{\mathtt{d}}$, and the diffusion speed $c_{\mathtt{d}}$ have to respect the following CFL conditions (See Courant *et al.* (1928) and Weisstein (2014)):

$$\omega_{\mathtt{d}} \leq \frac{\Delta x}{\Delta t} \qquad \textit{and} \qquad c \leq \frac{\Delta x^2}{\Delta t}, \tag{1}$$

where Δx corresponds to the discretization space step and Δt corresponds to the discretization time step.

Therefore we can conclude that, whatever the value of the other parameters :

$$\omega_{\mathtt{d}} \left[0; \tfrac{\Delta x}{\Delta t}\right]$$

$$c \in \left[0; \tfrac{\Delta x^2}{\Delta t}\right]$$

In the first exploration, we decided to parametrize the mesh with a time step of 0.001 and a space step of 0.025. Therefore:

$$\omega_{\mathtt{d}} \leq \frac{\Delta x}{\Delta t} = \frac{0.025}{0.001} = 25 \qquad \text{and} \qquad c \leq \tfrac{\Delta x^2}{\Delta t} = \tfrac{0.025^2}{0.001} = 0.625 \tag{2}$$

Hence, we have :

$$\omega_{\mathtt{d}} \in [0; 25]$$
$$\mathrm{c} \in [0; 0.625]$$

4. Saturation of the impact of the other parameters on the model

A saturation effect of the impact of the parameters $r_{\mathtt{d}}$, $f_{\mathtt{d}}$ and $u_{\mathtt{d}}$ on the model has already been introduced in Flourent *et al.* (Submitted). Indeed the range of values of each of these parameters corresponds to an interval from 0 to the saturation level of the impact of this parameter on the model.

To know for each parameter the value of this saturation level, we calculated an indicator $I_{ijk}(p_n, p_{n+1})$, corresponding to the evolution speed of the *Output Curve* profile according to the value of the studied parameter, p (where p can be $r_{\mathtt{d}}$, $f_{\mathtt{d}}$ or $u_{\mathtt{d}}$).

In the first studies we already fixed $c_{\mathtt{d}}$ at 0.001. Hence, to study the saturation of a given parameter p, we took the the three other ones among $\omega_{\mathtt{d}}$, $r_{\mathtt{d}}$, $f_{\mathtt{d}}$ and $u_{\mathtt{d}}$, except p. Then, we generated m curves from m increasing values p_n, $n \in [\![1, n]\!]$, taken in a wide range of values and we calculated:

$$I_{ijk}(p_n, p_{n+1}) = \frac{\left\| O_{ijk}^{n+1} - O_{ijk}^{n} \right\|}{\left\| O_{ijk}^{n+1} \right\|}, \tag{3}$$

for every i, j, k in their range. In 3,
$\|f\| = \int_0^1 |f(t)|\, dt$
n corresponds to the rank of the studied parameter p,
i corresponds to the rank of the first parameter,
j corresponds to the rank of the second parameter,
k corresponds to the rank of the third parameter,
O_{ijk}^{n} corresponds to the *Output Curve* generated from the values i, j, k, n of the

parameters.
And we imposed,

$$i \neq p,\ i \neq j,\ i \neq k,\ j \neq p,\ j \neq k \ and\ k \neq p \tag{4}$$

For example, to study the evolution of the impact of the parameter $u_{\mathtt{d}}$ we calculated $I_{ijk}(u_{\mathtt{d},n}, u_{\mathtt{d},n+1})$ for several values. Hence, $I_{ijk}(u_{\mathtt{d},n}, u_{\mathtt{d},n+1})$ corresponds to a relative difference between two curves generated from two consecutive values of $u_{\mathtt{d}}$: $u_{\mathtt{d},n}$ and $u_{\mathtt{d},n+1}$. In this case, i corresponds to the value of $\omega_{\mathtt{d}}$, j corresponds to the value of $r_{\mathtt{d}}$ and k corresponds to the value of $f_{\mathtt{d}}$.

For a given values of $r_{\mathtt{d}}$, $\omega_{\mathtt{d}}$ and $f_{\mathtt{d}}$, we calculated $I_{ijk}(u_{\mathtt{d},n}, u_{\mathtt{d},n+1})$ for 15 values of $u_{\mathtt{d}}$. We obtained the following curve:

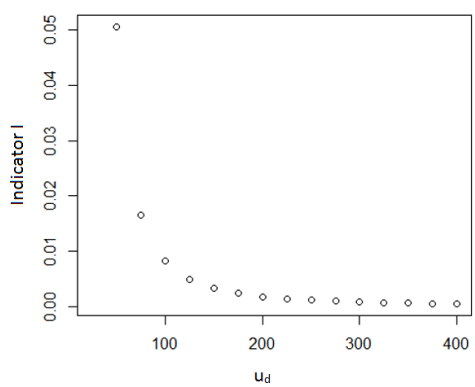

Figure 2: The value of the indicator I according to the value of $u_{\mathtt{d}}$.

In this case, the upper bound of $u_{\mathtt{d}}$ ($u_{\mathtt{d}\mathrm{Upp}}^{ijk}$), corresponds to the value from which the distance between two consecutive curves is lower than 0.005. So, $u_{\mathtt{d}\mathrm{Upp}}^{ijk}$ corresponds to the lowest value of $u_{\mathtt{d},n+1}$ such as $I_{ijk}(u_{\mathtt{d},n}, u_{\mathtt{d},n+1}) \leq 0.005$.

Therefore, we calculated p_{Upp}^{ijk} corresponding to the upper bound of the parameter under study, p, for p corresponding to $r_{\mathtt{d}}$, $f_{\mathtt{d}}$ or $u_{\mathtt{d}}$. Then we constructed a table in the following form :

i	j	k	p_{Upp}^{ijk}

equivalent to Table 1 corresponding to a Design of Experiments.

5. Determination of the ranges of values of $u_{\mathtt{d}}$, $f_{\mathtt{d}}$ and $r_{\mathtt{d}}$

5.1. The range of values of $u_{\mathtt{d}}$

In order to find the upper bound of $u_{\mathtt{d}}$ whatever the value of $f_{\mathtt{d}}$, $\omega_{\mathtt{d}}$ and $r_{\mathtt{d}}$, we calculated $u_{\mathtt{d}\mathrm{Upp}}^{ijk}$ for several values of $f_{\mathtt{d}}$, $\omega_{\mathtt{d}}$ and $r_{\mathtt{d}}$. Hence, the upper bound of $u_{\mathtt{d}}$ corresponds to the higher obtained value of $u_{\mathtt{d}\mathrm{Upp}}^{ijk}$.

Table 1: Design of Experiments to determine the upper bound of each parameter whatever the values of the others.

i	j	k	p_{Upp}^{ijk}
i_1	j_1	k_1	.
i_1	j_1	k_2	.
.	.	.	.
.	j_1	k_n	.
.	.	.	.
.	.	.	.
.	j_n	k_1	.
.	j_n	k_2	.
.	.	.	.
i_1	j_n	k_n	.
.	.	.	.
i_n	j_1	k_1	.
i_n	j_1	k_2	.
.	.	.	.
.	j_1	k_n	.
.	.	.	.
.	.	.	.
.	j_n	k_1	.
.	j_n	k_2	.
.	.	.	.
i_n	j_n	k_n	.

We tested 10 values of each parameter, taken between 0 and 25 for $\omega_{\mathtt{d}}$ and between 0 and 400 for $f_{\mathtt{d}}$ and $r_{\mathtt{d}}$. For each tested triplet $(f_{\mathtt{d}}, \omega_{\mathtt{d}}, r_{\mathtt{d}})$ we calculated the value of $u_{\mathtt{d}_{\mathrm{Upp}}}^{ijk}$ (Figure 3).

We noticed that $u_{\mathtt{d}_{\mathrm{Upp}}}^{ijk}$ does not exceed 200, whatever the value of $f_{\mathtt{d}}$, $\omega_{\mathtt{d}}$ and $r_{\mathtt{d}}$. Therefore, we can conclude that the upper bound of $u_{\mathtt{d}}$ is worth 200. Hence, $u_{\mathtt{d}}$ always belongs to the interval $[0; 200]$.

5.2. The range of values of $f_{\mathtt{d}}$

As, previously, to find the upper bound of $f_{\mathtt{d}}$ whatever the value of $u_{\mathtt{d}}$, $\omega_{\mathtt{d}}$ and $r_{\mathtt{d}}$, we calculated the value of $f_{\mathtt{d}_{\mathrm{Upp}}}^{ijk}$ for several values of $u_{\mathtt{d}}$, $\omega_{\mathtt{d}}$ and $r_{\mathtt{d}}$. We tested 10 values of each parameter, taken between 0 and 25 for $\omega_{\mathtt{d}}$ and between 0 and 400 for $u_{\mathtt{d}}$ and $r_{\mathtt{d}}$ and for each tested triplet $(u_{\mathtt{d}}, \omega_{\mathtt{d}}, r_{\mathtt{d}})$ we calculated the value of $f_{\mathtt{d}_{\mathrm{Upp}}}^{ijk}$.

In the left picture of Figure 4 the range of values reached by $f_{\mathtt{d}_{\mathrm{Upp}}}^{ijk}$ is given for every value of $i = u_{\mathtt{d}}$. The middle and the right pictures show the same for every value of $j = r_{\mathtt{d}}$ and $k = \omega_{\mathtt{d}}$.

The left picture of Figure 4 illustrates that $f_{\mathtt{d}_{\mathrm{Upp}}}^{ijk}$ does not depend on the value of $u_{\mathtt{d}}$. Yet, looking at the two other pictures, it seems there exists a relation between the values reached by $f_{\mathtt{d}_{\mathrm{Upp}}}^{ijk}$ and $\omega_{\mathtt{d}}$ and between the values reached by $f_{\mathtt{d}_{\mathrm{Upp}}}^{ijk}$ and $r_{\mathtt{d}}$.

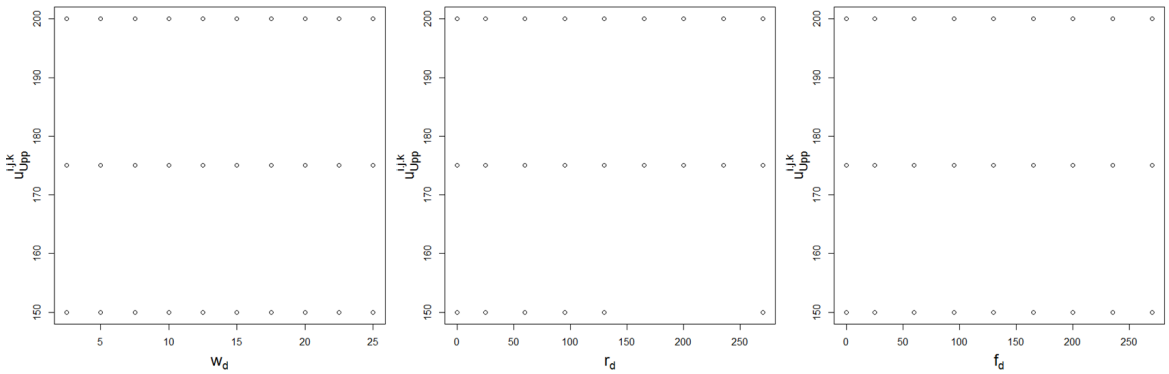

Figure 3: The value of $u^{ijk}_{\mathtt{d}\mathrm{Upp}}$ according to the values of $\omega_{\mathtt{d}}$, $r_{\mathtt{d}}$ and $f_{\mathtt{d}}$.

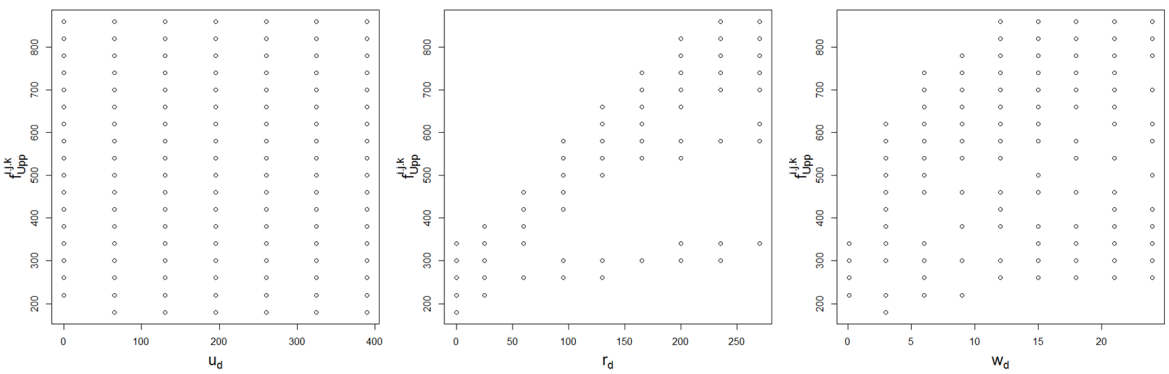

Figure 4: The value of the indicator $f^{ijk}_{\mathtt{d}\mathrm{Upp}}$ according to the values of $u_{\mathtt{d}}$, $r_{\mathtt{d}}$ and $\omega_{\mathtt{d}}$.

Figure 5 shows that the relation between the range of values of $f^{ijk}_{\mathtt{d}\mathrm{Upp}}$ and $r_{\mathtt{d}}$ is stronger. Therefore we studied more precisely this relation.

The highest value of $f^{ijk}_{\mathtt{d}\mathrm{Upp}}$ is obtained for $\omega_{\mathtt{d}} = 15$ (Figure 4). Hence, we fixed $\omega_{\mathtt{d}}$ at this values and we fix $u_{\mathtt{d}}$ at an average value of 100. Then, we calculated $f^{ijk}_{\mathtt{d}\mathrm{Upp}}$ for several values of $r_{\mathtt{d}}$ (Figure 6).

The curve in Figure 6 can be modeled via an equation in the form:

$$f_{\mathtt{d}\mathrm{Upp}} = \frac{A}{1 + B \cdot \exp(-C \cdot r_{\mathtt{d}}^{D})} + E \tag{5}$$

To fit the parameters A, B, C, D and E of Equation 5 on the curve of Figure 6, we used the function *directL* developed by Johnson (2008), which is embedded in R (R Core Team (2014)) and applying the DIRECT algorithm developed by Finkel (2003). After the fitting we obtained the results contained in Table 2.

The objective of this approach is to find the upper bound of $f_{\mathtt{d}}$ whatever the value of $r_{\mathtt{d}}$. Therefore we majored the value of E in order to be sure that the range of values of $f_{\mathtt{d}}$ includes the maximum value of $f_{\mathtt{d}}$. Hence, we fixed $E = 345$ (Figure 7).

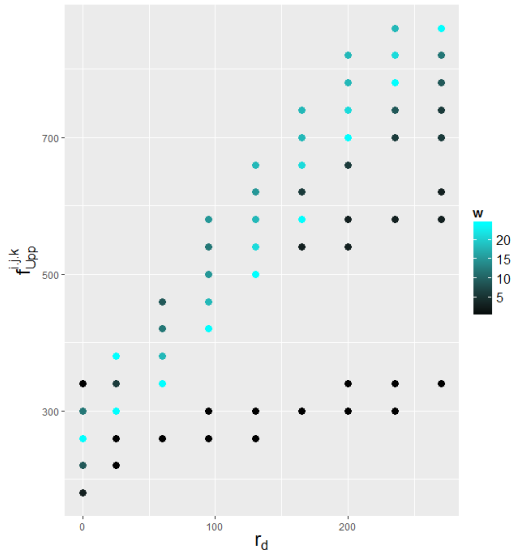

Figure 5: The value of the indicator f_{Upp} according to the value of $r_{\mathtt{d}}$ colored according to $\omega_{\mathtt{d}}$.

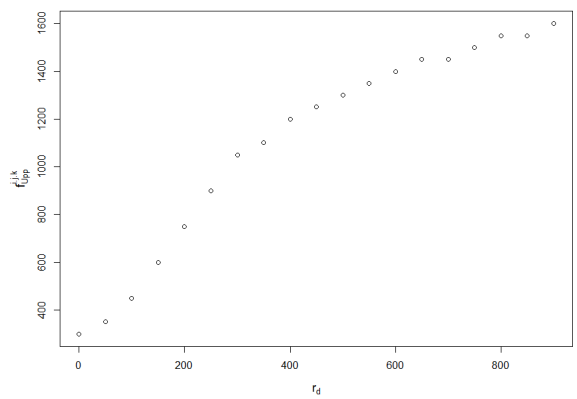

Figure 6: The value of the indicator $f_{\mathtt{d}\mathrm{Upp}}^{ijk}$ according to the value of $r_{\mathtt{d}}$.

Table 2: Coefficient of the equation linking $f_{\mathtt{d}\mathrm{Upp}}^{ijk}$ and $r_{\mathtt{d}}$.

Parameter	Value
A	1394.7
B	1853.1
C	1.83
D	0.25
E	272.3

5.3. The range of values of $r_{\mathtt{d}}$

While doing the early exploration, we observed that the coefficient $r_{\mathtt{d}}$ can take a large range of values before reaching a saturation level. Yet, as we saw in Flourent *et al.* (Submitted), the delay $r_{\mathtt{d}}$, may be offset by the convection speed $\omega_{\mathtt{d}}$. Indeed, a low convection speed associated to a low delay may induce the same kinetics as a high convection speed associated to an important delay. Hence, a couple $(\omega_{\mathtt{d},H}; r_{\mathtt{d},H})$ of high values of $\omega_{\mathtt{d}}$ and $r_{\mathtt{d}}$ can be equivalent to a couple $(\omega_{\mathtt{d},L}; r_{\mathtt{d},L})$ of lower values. Thus, this

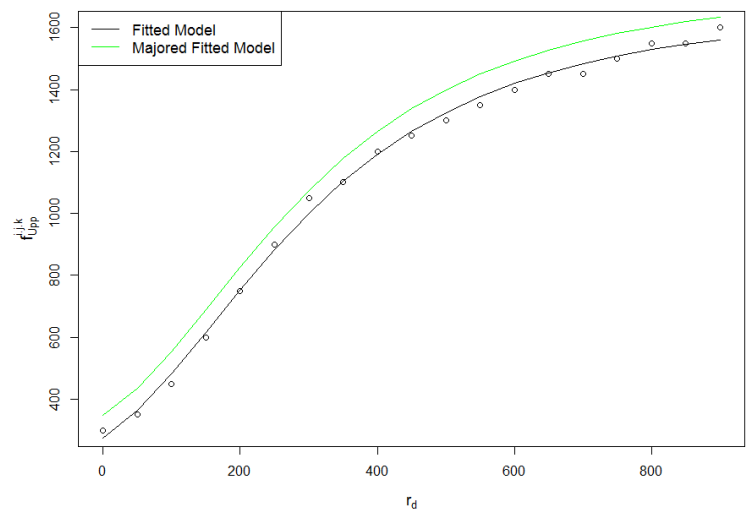

Figure 7: The relation linking $r_{\mathtt{d}}$ and $f_{\mathtt{d}\mathrm{Upp}}^{ijk}$

compensation effect permits to limit the range of values of $r_{\mathtt{d}}$ to a reduced interval but ensuring the capability of the model to simulate all the kinetics which can be observed.

We made the assumption that the highest value reachable by $r_{\mathtt{d}}$ is 900. Hence, we fixed $r = 900$. Then for several couples $(\omega_{\mathtt{d},i}; r_{\mathtt{d},i})$, where $r_{\mathtt{d},i} = 900$ and $\omega_{\mathtt{d},i} \in [2; 25]$, we sought the couple $(\omega_{\mathtt{d},t}; r_{\mathtt{d},t})$ permitting to generate the same *Output Curve* but from lower values of $\omega_{\mathtt{d}}$ and $r_{\mathtt{d}}$.

To do that we fixed $u_{\mathtt{d}} = 200$ and $f_{\mathtt{d}} = 900$, high values of these parameters. We selected 20 values of $\omega_{\mathtt{d},t} \in [0.2; 5]$. Then, for each couple $(\omega_{\mathtt{d},i}; r_{\mathtt{d},i})$ and each value of $\omega_{\mathtt{d},t}$, we sought the value of $r_{\mathtt{d}}$ minimizing the Relative Residual Sum of Squares ($RRSS$), measuring the difference between the curve generated from the couple $(\omega_{\mathtt{d},i}; r_{\mathtt{d},i})$ and the curve generated from the couple $(\omega_{\mathtt{d},t}; r_{\mathtt{d}t})$:

$$RRSS((\omega_{\mathtt{d},i}; r_{\mathtt{d},i}); (\omega_{\mathtt{d},t}; r_{\mathtt{d}t})) = \left(\sum_{k=1}^{n} \left(\frac{O(\omega_{\mathtt{d},i}; r_{\mathtt{d},i}) - O(\omega_{\mathtt{d},t}; r_{\mathtt{d},t})}{O(\omega_{\mathtt{d},i}; r_{\mathtt{d},i})} \right)^2 \right), \tag{6}$$

where $O(\omega_{\mathtt{d},k}; r_{\mathtt{d},k})$ corresponds to the *Output Curve* generated from the couple $(\omega_{\mathtt{d},k}; r_{\mathtt{d},k})$ and n is the number of points of the *Output Curves*.

To determined the value $r_{\mathtt{d},t}$ we used the optimization algorithm DIRECT. This algorithm determines the value of $r_{\mathtt{d},t}$ by minimizing Indicator (6).

At the end of this step, we obtained 20 couples $(\omega_{\mathtt{d},t}; r_{\mathtt{d},t})$ for each couple $(\omega_{\mathtt{d},i}; r_{\mathtt{d},i})$. For each couple $(\omega_{\mathtt{d},i}; r_{\mathtt{d},i})$ we selected among the 20 obtained couples $(\omega_{\mathtt{d},t}; r_{\mathtt{d},t})$, the couple associated to the lowest $RRSS$. We finally obtained Table 3, in which R^2 corresponds to the squared correlation between the *Output Curves* $O(\omega_{\mathtt{d},i}; r_{\mathtt{d},i})$ and $O(\omega_{\mathtt{d},t}; r_{\mathtt{d},t})$.

Therefore, for each couple $(\omega_{\mathtt{d},i}; r_{\mathtt{d},i})$ we found an equivalent couple $(\omega_{\mathtt{d},t}; r_{\mathtt{d},t})$ presenting a lower value of $r_{\mathtt{d}}$.

Table 3: Coefficient of the equation linking $f_{\mathtt{dUpp}}$ and $r_{\mathtt{d}}$.

$\omega_{\mathtt{d},i}$	$r_{\mathtt{d},i}$	$\omega_{\mathtt{d},t}$	$r_{\mathtt{d},t}$	$RRSS$	R^2
2	900	0.7	284	35.2	0.98
4.5	900	1	194.4	53.9	1
7	900	1.5	185.1	61.2	1
9.5	900	2	177.5	76.8	1
12	900	2.5	171.4	76.6	1
14.5	900	3	166.6	66.2	1
17	900	3.5	162.9	55.9	1
19.5	900	4	160.1	48.7	1
22	900	4.5	157.9	44.2	1
24.5	900	5	156.1	42.1	1

The couple $(\omega_{\mathtt{d},t}; r_{\mathtt{d},t})$ equivalent to the couple $(\omega_{\mathtt{d},i} = 2; r_{\mathtt{d},i} = 900)$ presents the highest value of $r_{\mathtt{d}}$, which is 284. Therefore, whatever the value of $f_{\mathtt{d}}$, $u_{\mathtt{d}}$ and $\omega_{\mathtt{d}}$ the upper bound of $r_{\mathtt{d}}$ is worth 284.

5.4. Deduction about the range of values of $f_{\mathtt{d}}$

We saw previously that the upper bound of $f_{\mathtt{d}}$ depends on the upper bound of $r_{\mathtt{d}}$. We demonstrated in the previous Section that the highest value of $r_{\mathtt{d}}$ is worth 284. Hence, from Equation 5, we deduced that the upper bound of $f_{\mathtt{d}}$ is worth 1035.

6. Conclusion about the ranges of values of the different parameters

To conclude, we can say that the ranges of values of $\omega_{\mathtt{d}}$ and $c_{\mathtt{d}}$, are imposed by the CFL conditions.

The upper bound of the parameter $u_{\mathtt{d}}$ was determined by studying the saturation of the effect of this parameter on the model.

A saturation of the impact of $f_{\mathtt{d}}$ on the model is also observed. Nevertheless, the saturation level of this parameter depends on the value of $r_{\mathtt{d}}$. Therefore we established a relation linking those two parameters.

The parameter $r_{\mathtt{d}}$ can take high values. But the existence of a compensation effect between $r_{\mathtt{d}}$ and $\omega_{\mathtt{d}}$ permits to limit the range of values of $r_{\mathtt{d}}$ to a reduced interval but ensuring the capability of the model to simulate all the kinetics which can be observed.

The determination of the value of the upper bound of $r_{\mathtt{d}}$ permitted then to deduce the upper bound of $f_{\mathtt{d}}$ by using the relation previously established.

Finally we obtained Table 4.

Table 4: The ranges of values of the different parameters.

Parameter	Range of values
ω	$[0; 25]$
c	$[0; 0.625]$
u	$[0; 200]$
r	$[0; 284]$
f	$[0; 1035]$

References

Courant, R., K. Friedrichs, and H. Lewy (1928). "Über die partiellen Differenzengleichungen der mathematischen Physik". In: *Mathematische annalen* 100.1, pp. 32–74.

Enright, W. (1989). "The Numerical Analysis of Ordinary Differential Equations: Runge Kutta and General Linear Methods". In: *SIAM Review* 31.4, pp. 693–693. DOI: 10.1137/1031147. eprint: https://doi.org/10.1137/1031147. URL: https://doi.org/10.1137/1031147.

Finkel, D. E. (2003). *DIRECT Optimization Algorithm*. North Carolina State University.

Flourent, H., E. Frénod, and V. Sincholle (Submitted). "An innovating Statistical Learning Tool based on Partial Differential Equations, intending livestock Data Assimilation".

Johnson, S. G. (2008). "The NLopt nonlinear-optimization package".

R Core Team (2014). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing. Vienna, Austria. URL: http://www.R-project.org/.

Soetaert, K., T. Petzoldt, and R. Woodrow Setzer (2010). "Solving Differential Equations in R: Package deSolve". In: *Journal of Statistical Software* 33.9, pp. 1–25. ISSN: 1548-7660. DOI: 10.18637/jss.v033.i09. URL: http://www.jstatsoft.org/v33/i09.

Weisstein, E. W. (2014). "Courant-friedrichs-lewy condition". In: *Wolfram MathWorld–A Wolfram Web Resource*.