



**HAL**  
open science

# Optimal spatial monitoring of populations described by reaction–diffusion models

Nicolas Parisey, Melen Leclerc, Katarzyna Adamczyk-Chauvat

► **To cite this version:**

Nicolas Parisey, Melen Leclerc, Katarzyna Adamczyk-Chauvat. Optimal spatial monitoring of populations described by reaction–diffusion models. *Journal of Theoretical Biology*, 2022, 534, pp.110976. 10.1016/j.jtbi.2021.110976 . hal-03651840

**HAL Id: hal-03651840**

**<https://hal.inrae.fr/hal-03651840>**

Submitted on 8 Jan 2024

**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.



Distributed under a Creative Commons Attribution - NonCommercial - NoDerivatives 4.0 International License

# Optimal spatial monitoring of populations described by reaction-diffusion models

Nicolas Parisey<sup>a,\*</sup>, Melen Leclerc<sup>a,1</sup> and Katarzyna Adamczyk-Chauvat<sup>b,1</sup>

<sup>a</sup>UMR 1349 IGEPP, INRAE, Le Rheu, 35653, France

<sup>c</sup>UR 1404 MaIAGE, INRAE, Jouy-en-Josas, 78352, France

## ARTICLE INFO

## Abstract

### Keywords:

optimal design of experiments  
sampling effort  
spatial population model  
optimal survey  
mechanistic-statistical modelling


Using spatialised population measurements and related geographic habitat data, it is feasible nowadays to derive parsimonious spatially explicit population models and to carry on their parameter estimation. To achieve such goal, reaction-diffusion models are common in conservation biology and agricultural plant health where they are used, for example, for landscape planning or epidemiological surveillance. Unfortunately, if the mathematical methods and computational power are readily available, biological measurements are not. Despite the high throughput of some habitat related remote sensors, the experimental cost of biological measurements are one of the worst bottleneck against a widespread usage of reaction-diffusion models. Hence we will recall some classical methods for optimal experimental design that we deem useful to spatial ecologist. Using two case studies, one in landscape ecology and one in conservation biology, we will show how to construct *a priori* experimental design minimizing variance of parameter estimates, enabling optimal experimental setup under constraints.

## 1. Introduction

Both empirical and theoretical studies have well established that population spread is an essential ecological process for understanding most of the observed population dynamics. Consequently, over the last two decades spatial ecology has become central either in theoretical ecology or in empirical approaches. Hence, population management for their control, capture or protection, has started to be considered at large spatial scales that match the inherent dispersal capacity of the considered species. Populations can be identified and monitored using various strategies (e.g. tracking, trapping). However, despite the development of new technologies and devices for ecological monitoring (e.g. imaging sensors for animal detection (Weinstein, 2018), autonomous Unmanned Aerial Vehicle (Cliff, Saunders & Fitch, 2018)), the survey of many species on large areas remains challenging, costly and empirically guided.

The growing interest for space in ecology has been accompanied by a proliferation of models and statistical methods for analyzing and predicting the spatial distribution of populations. Nowadays, existing statistical methods allows one to infer the dynamics of spreading populations from noisy and sparse data. However, statistical inference of spatio-temporal models from common monitoring data can be subject to practical identifiability issues and is often associated

\*Corresponding author

 [nicolas.parisey@inrae.fr](mailto:nicolas.parisey@inrae.fr) (N. Parisey); [melen.leclerc@inrae.fr](mailto:melen.leclerc@inrae.fr) (M. Leclerc); [katarzyna.adamczyk@inrae.fr](mailto:katarzyna.adamczyk@inrae.fr) (K. Adamczyk-Chauvat)

ORCID(s): 0000-0003-2439-3809 (N. Parisey); 0000-0002-5314-461X (M. Leclerc)

<sup>1</sup>Both authors contributed equally to this work.

36 with an important uncertainty on parameters estimates. Thus, given the inherent difficulty of collecting spatial data on  
 37 most population, model-based forecasts are still associated with an important uncertainty, which could theoretically  
 38 be reduced by more dense and efficient monitoring strategies.

39 When planning a survey of population over a large area, one has to make some choices giving financial and human  
 40 constraints: which observational method (transects, traps), where and when to start the survey, the number and times  
 41 for repeating the survey. These questions can either be addressed empirically or using model-based techniques that  
 42 can be organized into three major groups of methods (Ucinski, 2004). The first group transform the problem into a  
 43 state-estimation one by augmenting the state vector (Malebranche, 1988). The second group is based on random fields  
 44 theory (Sun, 1999) whereas the third group corresponds to the classical theory of the design of experiments (Ucinski,  
 45 2004). We can also point out a fourth group that consists in adopting an heuristic to find the best sampling strategies  
 46 among tested situations (Bellot, Poggi, Baudry, Bourhis & Parisey, 2018). Based on either of these groups, methods  
 47 for designing ecological sampling has already been addressed by several authors (Hooten, Wikle, Sheriff & Rushin,  
 48 2009; Williams, Hooten, Womble & Bower, 2018) but it has not yet percolated in the modellers community and is still  
 49 seldom considered in population ecology.

50 The use of optimal design of experiments is perhaps one of the lesser known (Cook, Gibson & Gilligan, 2008)  
 51 whereas it is well developed for optimal sensor placements in spatially-distributed systems with non-linear dynamics  
 52 models (Ucinski, 2004). It corresponds to an entire branch of statistics that provides criteria measuring the amount of  
 53 information about unknown model parameters carried out by the observed data.

54 Here, "experiments" is taken *sensus lacto* and refers to controlled observations of populations within their living  
 55 areas that are planned given the constraints of available resources and with subsequent statistical analysis in mind.

56 In this study we use the optimal design of experiments to tackle the question of ecological monitoring of populations  
 57 described by reaction-diffusion models. This type of partial differential equations originally emerged in chemistry  
 58 for analyzing the change in space and time of chemical substances before becoming one of the most important  
 59 mathematical model for the study of spreading dynamical processes in biology and ecology. Here, we consider reaction-  
 60 diffusion equations that provides a parsimonious description of a spreading population in a domain  $\Omega$  included in  $\mathbb{R}^2$

61 :

$$\frac{\partial u(x, t)}{\partial t} = D(x)\Delta u(x, t) + f(u, x, t) \quad (1)$$

62 where  $u(x, t)$  is the population density at time  $t$  and location  $x \in \Omega$ ,  $D(x)$  corresponds to the dispersion rate at  $x$ ,  $\Delta u(x, t)$   
 63 is the Laplace operator of  $u$  evaluated at  $(x, t)$  that describes the random movement of individuals, and  $f(u, x, t)$  is a

64 general reaction function accounting for local birth, death and interactions within the spreading population. To fully  
65 describe the system, one needs to define some boundary conditions on  $\Omega$  as well as initial conditions  $u_0(x) = u(0, x)$ .

66 In addition to their ability to capture the essential observed patterns of spreading populations, reaction-diffusion  
67 systems have been well studied by mathematicians for a long time. Thus, though the mathematical and the numerical  
68 analysis of some reaction-diffusion equations still challenge mathematicians, several systems are now well character-  
69 ized and understood. The most popular one is the Fisher-KPP system, that applies in ecology, for which  $D(x) = D$  and  
70 the reaction term accounts for a limiting carrying capacity of the environment  $f(u, x, t) = u(1-u)$ . Parameter estimation  
71 of mechanistic PDE models from noisy observational data, obtained during common population monitoring strategies,  
72 can be performed using different statistical methods. In any case, one needs to link the mechanistic PDE model that  
73 describes the changes in space and time of the population density with a probabilistic model describing the observation  
74 process. This approach refers to the physical-statistical or mechanistic-statistical models that are specific instances of  
75 the more general hierarchical state-space model framework (Clark & Bjørnstad, 2004; Soubeyrand & Roques, 2014;  
76 Wikle, 2003). Such approach has been successfully used for inferring spatio-temporal processes on large areas for  
77 invasive and beneficial insects (Parisey, Bourhis, Roques, Soubeyrand, Ricci & Poggi, 2016; Roques, Soubeyrand  
78 & Rousselet, 2011; Roques & Bonnefon, 2016), plant diseases (Abboud, Bonnefon, Parent & Soubeyrand, 2019) or  
79 aquatic and terrestrial mammal species (Louvrier, Papaïx, Duchamp & Gimenez, 2020). As illustrated by Williams et al.  
80 (2018), the optimal design of experiments can be applied on mechanistic-statistical systems for ecological monitoring,  
81 but is yet poorly considered by both modellers and ecologists collecting spatio-temporal data.

82 In this article we first introduce the main concepts of the optimal design of experiments applied on mechanistic-  
83 statistical systems for population dynamics. Then, we consider two example populations whose spatio-temporal  
84 dynamics is described by reaction-diffusion equations, but monitored with different strategies: beneficial insects for  
85 agriculture that are monitored using Barber pitfall traps at various locations within the landscape (Parisey et al., 2016),  
86 and invasive wild horses that are counted during aircraft transects (Beeton & Johnson, 2019). Assuming that the more  
87 realistic degree of freedom in the existing monitoring designs is the choice of the locations of observations we consider  
88 static designs, where locations of sampled populations are fixed before the start of the survey and don't change in time,  
89 and focus on the spatial aspect of monitoring. We use D-optimal designs that define optimal locations for population  
90 observation and show how this framework can help to design ecological monitoring strategies over large areas under  
91 identical constraints. We finish the paper by discussing the use of the optimal design of experiments framework to  
92 improve the monitoring of populations over large areas, especially to reduce the cost of sampling and support the link  
93 between modelling and empirical studies.

## 94 2. Mechanistic-statistical model for population dynamics

95 We consider a mechanistic-statistical approach that combines a reaction-diffusion system which describes essential  
 96 spatio-temporal dynamics of a population with an observational equation describing the stochastic process leading to  
 97 detection and enumeration of individuals at location  $x \in \Omega$  and time  $t$  (as in Soubeyrand & Roques (2014)). However,  
 98 following the framework used in control system analysis or optimal design for parameter identification (Ucinski,  
 99 2004), we decompose the observation process into two parts: 1) the measurement process which links the ecological  
 100 process and factors that affect the measurement of population density (e.g. equipment features, operator behaviour,  
 101 environmental effect), and 2) the data process that links population data with the measurement process. Then, given  
 102 some boundary conditions on  $\Omega$  the spatially distributed system is described by the following hierarchical system :

$$\left\{ \begin{array}{l} u(x, 0) = u_0(x) \\ \frac{\partial u(x, t)}{\partial t} = D\Delta u(x, t) + f(u, x, t, \theta_p) \\ \eta(x, t) = g(u(x, t), \theta_m) \\ y \sim \mathcal{D}(\eta(x, t), \theta_d) \end{array} \right. \quad (2)$$

103 where  $\eta(x, t)$  is the measurement process determined from population density  $u(x, t)$  with function  $g(\cdot)$ , and the data  
 104 process corresponds to the draw of population data  $y$  from a general probability distribution  $\mathcal{D}$ . Depending on the  
 105 considered system, parameters that needs to be estimated, or identified, from the monitoring data  $y$  can occur in the  
 106 population process  $(D, \theta_p)$ , the measurement process  $(\theta_m)$  and the data process  $(\theta_d)$ . The set of unknown parameters is  
 107 thus given by  $\theta = (D, \theta_p, \theta_m, \theta_d)$ . Albeit initial conditions  $u_0(x)$  of reaction-diffusion systems may also be estimated  
 108 from noisy data (Abboud et al., 2019; Soubeyrand & Roques, 2014), in this study we assume that  $u_0(x)$  is known  
 109 and fixed. Parameter estimation can be achieved using maximum likelihood estimation (see next section). It generally  
 110 involves a numerical optimization that requires the simulation of the reaction-diffusion system with a suitable numerical  
 111 method (Fornberg & Sloan, 1994; Hundsdorfer & Verwer, 2003).

## 112 3. Optimal design

113 For monitoring a population, the questions that we should address at first are where to look at the population  
 114 and when to observe the targeted population? Though the time scheduling of the monitoring is often constrained by  
 115 numerous factors, the choice of spatial locations generally offers more degree of freedom. In the following, we assume  
 116 that the dates of the monitoring design (or experiment *sensus lato*) are scheduled in advance and also that the sensors

117 locations will not change in time. This simplifies the problem of defining optimal monitoring designs and thus an  
 118 optimal spatial distribution of observational locations over the survey area.

119 This question can be addressed through experimental design theory, a branch of applied statistics introduced by  
 120 Sir Ronald Fisher in his seminal book edited in 1935 (Fisher, 1935). The purpose of this section is to introduce some  
 121 key-notions of experimental design theory, essential for understanding its application on the use cases presented in  
 122 section 4. For a complete description of the statistical framework the reader can refer to Silvey (2013), Atkinson,  
 123 Donev & Tobias (2007) or Walter & Pronzato (1997). We start by introducing the key notions of optimal design and  
 124 then we give their interpretation for the mechanistic-statistical model presented in equation 2. We indifferently use the  
 125 terminology inherited from the design of experiments theory. Therefore, the experiments will refer to the survey of the  
 126 populations and sensors corresponds to any monitoring device or a visual inspection.

127 Let us consider a random variable  $y$  with a probability distribution depending on: (1) a vector of real variables  $x$   
 128 that can be chosen by the experimenter, (2) and a vector of parameters  $\theta$ , supposed to be fixed and unknown for the  
 129 experimenter. We assume that  $x$  belongs to the set  $\mathcal{X} \subset \mathbb{R}^r$  and that  $\theta$  belongs to a parametric space  $\Theta \subset \mathbb{R}^p$ . The  
 130 variables composing  $x$  are called control variables. Let us suppose that for a given  $x$  and  $\theta$  the distribution of  $y$  is  
 131 given by a probability density function  $p(y|x, \theta)$ . The experimenter is allowed to take  $N$  independent observations on  
 132  $y$  at vectors  $x_1, \dots, x_N$  chosen from the set  $\mathcal{X}$ . The set  $\mathbf{x} = \{x_1 \dots x_N\}$  will be referred as  $N$ -observation design. The  
 133 question, primarily, is how to select the design  $\mathbf{x}$  ?

134 The criterion of choice depends on the purpose of the experiment. Here our primary interest is in estimating the  
 135 parameter  $\theta$  from the experimental data. Let us denote by  $\underline{x}$  a vector  $(x_1, \dots, x_N)$  and let  $\underline{y}$  be a vector of values of  $y$   
 136 taken at  $x_i$ . The log-likelihood function of  $\theta$  is defined as:

$$\log \Lambda(\theta; \underline{x}, \underline{y}) = \sum_{i=1}^N \log(p(y_i|x_i, \theta)) \quad (3)$$

The maximum likelihood estimate of  $\theta$  maximizes  $\log \Lambda$  over  $\Theta$ :

$$\hat{\theta}(\underline{x}, \underline{y}) = \arg \max_{\theta \in \Theta} \log \Lambda(\theta; \underline{x}, \underline{y})$$

Under some regularity conditions on the family of densities  $\{p_\theta : \theta \in \Theta\}$  the estimate  $\hat{\theta}$  is asymptotically normal as  
 the sample size  $N$  tends to infinity (see Lehmann & Casella (1998) for instance). Moreover, the asymptotic variance of  
 $\hat{\theta}$  reaches its lower bound given by Cramer-Rao inequality and is equal to the inverse of the Fisher Information Matrix,

defined as:

$$I(\theta; \underline{x}) = \sum_{i=1}^N \mu(x_i, \theta) = \sum_{i=1}^N \text{Var} \left( \frac{\partial}{\partial \theta} \log p(y|x_i, \theta) \right) \quad (4)$$

137 where  $\mu(x, \theta)$  is the information matrix for an observation on  $y$  taken at  $x$ . Roughly speaking an optimal design  $\mathbf{x}_\star$   
 138 makes the variance of  $\hat{\theta}$  “as small as possible” or, alternatively, makes the Fisher matrix “as large as possible”. To  
 139 be more precise we seek for an  $\mathbf{x}_\star$  that maximizes some real-valued function  $\phi$  of  $I(\theta; \underline{x})$ . We assume here that  $\phi$  is  
 140 homogeneous and concave.

We consider the design with  $n$  distinct vectors  $x_1, \dots, x_n$  replicated  $r_1, \dots, r_n$  times, where  $\sum_{i=1}^n r_i = N$ . We assign  
 to  $\mathbf{x}$  a discrete probability distribution  $\xi^n$  which puts the probabilities  $\omega_i = r_i/N$  at  $x_i$ . Consequently, the Information  
 Matrix (4) can be rewritten as :

$$I(\theta; \underline{x}) = N M(\theta; \xi^n) \text{ where } : M(\theta; \xi^n) = \sum_{i=1}^N \omega_i \mu(x_i, \theta) \quad (5)$$

141 As a criterion function  $\phi$  is homogeneous, optimizing  $\phi(I(\theta; \underline{x}))$  amounts to optimizing  $\phi(M(\theta; \xi^n))$  over  $\xi^n$ . An  
 142 optimal design  $\xi_\star^n$  is thus defined as:

$$\xi_\star^n(\theta) = \arg \max_{\xi^n \in \Xi} \phi(M(\theta; \xi^n)) \quad (6)$$

143 where the matrix  $M$  is assumed to be non-singular.

144 *D-optimal design* The choice of the optimality criterion relies on the final purpose of statistical analysis. The  
 145 experimenter may wish to improve the precision of  $\hat{\theta}$  or to reduce the variance of the predicted values of  $y$  or again  
 146 to better discriminate between the candidate models. The main optimality criteria and the corresponding functions  $\phi$   
 147 can be found for example in Walter & Pronzato (1997). Because of its versatility of purpose, one of the mostly used  
 148 criterion is the D criterion leading to a D-optimal design, maximizing the logarithm of the determinant of the matrix  
 149  $M$ :

$$\phi(M(\theta; \xi^n)) = \log \det M(\theta; \xi^n) \quad (7)$$

150 A D-optimal design aims at minimizing the volume of confidence ellipsoid for model parameters. Consequently  
 151 the statistical treatment of experimental results become more efficient. For example, it helps to assess if a model's

152 parameter, and its linked hypothesis, has a non negligible effect on a population dynamics e.g. whether a parameter  
 153 is different from zero. The D-optimal design also helps strengthen parameter comparison. For example it's usually  
 154 important to estimate which habitat is the best (or worst) for a given species e.g. to significantly rank their  
 155 corresponding growth rates. Finally, this design can also lead to minimizing the maximum variance of the predicted  
 156 values (Kiefer, 1974).

157 *Exact and approximate design.* When the set of the matrices defining a domain of (7) is discrete, the design is said  
 158 to be exact. The problem of finding an exact optimal design  $\xi_{\star}^n$  could be numerically challenging, especially for the  
 159 large values of  $N$ . The solution proposed in the theory of optimal design is to calculate the optimum of  $\phi$  over the  
 160 extended domain and to look for the discrete  $N$ -observational design which is “close” to the optimum. The definition  
 161 of the matrix  $M$  can be extended by considering the set of all probability distributions on  $\mathcal{X}$  in place of the discrete  
 162 distributions  $\xi^n$ :

$$M(\theta; \xi) = \int_{\mathcal{X}} \mu(x; \theta) \xi(dx) \quad (8)$$

where  $\xi$  is a probability distribution on  $\mathcal{X}$ . Then the approximate solution of the initial optimization problem can  
 be calculated. An optimal (continuous) approximate design  $\xi_{\star}$  is defined as:

$$\xi_{\star}(\theta) = \arg \max_{\xi \in \Xi} \phi(M(\theta; \xi))$$

163 for the matrix  $M$  given by the formula (8).

164 *Average and local optimal design.* An optimal design for a non-linear model depends on the unknown parameter  
 165 value and is referred to as a local design. Different methods are proposed to “remove” the dependence from  $\theta$ . One  
 166 possible approach is to assume that  $\theta$  is a random variable following a known distribution  $\pi$  and to maximize the  
 167 expected value of the criterion, calculated with respect to  $\pi$ . Consequently, an on-average exact design maximizes:

$$\mathbb{E}_{\theta} \phi(M(\theta; \xi^n)) = \int_{\Theta} \phi(M(\theta; \xi^n)) \pi(d\theta) \quad (9)$$

*Fisher Information Matrix for generalized regression model.* According to the definition given in Atkinson,  
 Fedorov, Herzberg & Zhang (2014), a generalized regression model is specified by the probability density  $p$  that



depends on  $x$  and  $\theta$  through the  $k$  dimensional regression function  $\eta(x, \theta)$ :

$$p(y|x, \theta) = p(y|\eta(x, \theta))$$

168 Consequently, the Fisher Information Matrix  $\mu(x, \theta)$  for a single observation on  $y$  can be rewritten as :

$$\begin{aligned} \mu(x, \theta) &= \text{Var} \left( \frac{\partial}{\partial \theta} \log p(y|\eta(x, \theta)) \right) \\ &= D_{\theta} \eta(x, \theta) \underbrace{\text{Var} \frac{\partial}{\partial \eta} \log p(y|\eta)}_{v(\eta)} D_{\theta}^T \eta(x, \theta) \end{aligned} \quad (10)$$

169 where  $D_{\theta} \eta(x, \theta)$  is a  $p \times k$  matrix of the first derivatives of the regression function with respect to  $\theta$  and  $v(\eta)$  is  
 170 a  $k \times k$  Fisher Information Matrix for the reparametrized model. The matrix  $v(\eta)$  is called the elemental information  
 171 matrix and is known for many useful probability densities.

172 *Design settings for spatial population monitoring.* We assume that the targeted population is observed at  $N$   
 173 locations  $x_1, \dots, x_N$ . For each location, the measurement is repeated  $T$ -times. The time measurements are fixed in  
 174 advance and common for all the locations whereas the set of the locations is solution to the optimal design problem.  
 175 The measurement taken at a location  $x_i$  at a time moment  $t$  is denoted by  $y(x_i, t)$ . Typically  $y(x_i, t)$  is either a population  
 176 count or its suitable transform. In terms of mechanico-statistical model introduced in section 2,  $y(x_i, t)$  is the result of a  
 177 data process, a random variable following a probability distribution  $\mathcal{D}$ . The distribution  $\mathcal{D}$  depends on a measurement  
 178 process  $\eta(x, \theta)$  governed by a partial differential equation and on a vector of parameters  $\theta$ , related to all the levels of the  
 179 hierachical model (see eq. 2). The likelihood of  $\theta$ , given the vector of observations  $(y(x_i, t))_{i=1, \dots, N}^{t=1, \dots, T}$  is given by equation  
 180 3. The optimal design problem adressed in this paper is to find an exact D-optimal on-average design  $\xi_*^n = (x_1, \dots, x_N)$   
 181 (implicitly  $\omega_1 = \dots = \omega_N = 1$ ). According to the short introduction given in this section,  $\xi_*^n$  satisfies :

$$\xi_*^n = \arg \max_{\xi_n} \int_{\mathcal{X}} \log \det(M(\theta, \xi^n)) \pi(d\theta) \quad (11)$$

182 where  $M(\theta, \xi^n)$  is calculated according to equation 5 and 10. From this point on, we assume that the parameter  $\theta$   
 183 follows a uniform distribution  $\mathcal{U}(\theta_{min}, \theta_{max})$ . In the following chapters, we illustrate the solution of optimal design  $\xi_*^n$   
 184 for two spatial monitoring examples.

## 185 4. Spatial population monitoring case studies

### 186 4.1. Agroecological case study

187 As a first example we consider the dynamics of *Poecilus cupreus*, a carabid beetle known for its weed seed  
 188 consumption, within agricultural landscapes. This beneficial insect often complete its life cycle in a couple of months  
 189 and is commonly monitored using pitfall traps that are picked up and replaced weekly. This species exhibits a single  
 190 peak during its activity season and is known to react differently to semi-natural habitats, grasslands and cereal fields  
 191 (Marrec, Badenhauer, Bretagnolle, Börger, Roncoroni, Guillon & Gauffre, 2015). This leads to a reaction-diffusion  
 192 model introduced in Parisey et al. (2016) with a birth rate decays parameter ( $\beta$ ), to mimick the single pick, and a  
 193 spatially heterogeneous growth rate  $r(x)$ , to express the habitat dependencies. In Parisey et al. (2016), the population  
 194 dynamics of carabids has been studied within a landscape of a few kilometer squared size. The dynamics of carabids  
 195 was described by the following mechanistic-statistical model:

$$\left\{ \begin{array}{l} u(x, 0) = u_0(x) \\ \frac{\partial u(x, t)}{\partial t} = D\Delta u(x, t) + (r(x)e^{-\beta t} - \mu)u(x, t) \\ \eta(x, t) = \zeta \int_{t-\tau}^t u(x, s) ds \\ y \sim \mathcal{P}(\eta(x, t)) \end{array} \right. \quad (12)$$

where  $u(x, t)$  is in this case the density of carabids,  $\beta$  describes the exponential speed at which the birth rate decays during the activity season,  $\mu$  is the death rate (i.e.  $\frac{1}{\text{life expectancy}}$ ),  $r(x)$  is an habitat specific growth rate such that :

$$r(x) = \left\{ \begin{array}{l} r_c \text{ if } x \text{ in a crop field} \\ r_s \text{ if } x \text{ in a semi-natural habitat} \\ r_g \text{ if } x \text{ in a grassland} \\ 0 \text{ otherwise (roads, buildings, ...)} \end{array} \right.$$

196 ,  $\zeta$  is a scaling factor that allows the integration of the carabids density over the surface of a non-attractive pitfall trap  
 197 from which count data is described by a Poissonian data process  $\mathcal{P}(\cdot)$  with intensity  $\eta(x, t)$ .

198 The population process is parametrized by  $(D, \theta_p)$  with  $\theta_p = (\beta, \mu, r_c, r_s, r_g)$  and the measurement process depends  
 199 on  $\theta_m = (\zeta)$ , that we considered known for a given type of traps. Hence the vector of unknown parameters were  
 200  $\theta = (D, \beta, \mu, r_c, r_h, r_g)$ . The parameters of the model were already estimated in Parisey et al. (2016) for an arbitrarily

201 chosen set of pitfall traps. We used these results in order to fix the bounds of parameters  $\theta$  in an on-average design.  
 202  $\theta_{min}$  and  $\theta_{max}$  were set so they were biologically relevant and coherent with previous estimations. All growth rates were  
 203 explored so they give half to twice as many descendants, during a season of several months, as previously estimated. The  
 204 diffusion coefficient varies between a population of individuals half to twice as fast. As  $\mu$  is linked with life expectancy,  
 205 it varies between a life twice as long and half as long. Finally,  $\beta$ , measuring birth decays during the activity season,  
 206 varies between 0.9 and 1.1 times its nominal value. Details for all parameters values can be found in appendix A.1.

207 In this example, as described in Parisey et al. (2016), the experiment consisted in placing  $N = 24$  pitfall traps in  
 208 an agricultural landscape, owned by several farms, and sample them over the course of a season, in that case  $T = 9$   
 209 times over the course of three months in 2010. The sampling design of this study was decided beforehand so that, first,  
 210 an agricultural field was chosen then several traps (here three) were planted, spatially grouped, in the field. From now  
 211 on, we will refer to this as a clustered design.

## 212 4.2. Conservation biology case study

213 As a second example we considered the dynamics of *Equus caballus*, an invasive feral horse that has detrimental  
 214 effects to the ecosystem of the Australian Alps and whose management through aerial culling or trapping and mustering  
 215 is under debate. This question has been addressed by Beeton & Johnson (2019) who used a two years aerial survey of  
 216 feral horse populations, consisting in line transect data within the Australia Alps region (Cairns, 2014). They derived  
 217 a reaction-diffusion model assuming horse populations are limited by density dependence in births via a logistic  
 218 model, and that their movement through the landscape depends on the local horse density and, spatially heterogeneous,  
 219 carrying capacity. We supplemented their spatio-temporal dynamics with a measurement process and a data process to  
 220 obtain a full mechanistic-statistical model. The measurement process model the use of aerial transects to assess horse  
 221 populations while the data process assumes normally distributed residuals with zero mean and variance  $\sigma^2$ . This lead  
 222 to a mechanistic-statistical system given by :

$$\left\{ \begin{array}{l} u(x, 0) = u_0(x) \\ \frac{\partial u(x, t)}{\partial t} = D\Delta \frac{u(x, t)}{K(x)} + (b(1 - \frac{u(x, t)}{K(x)}) - \mu)u(x, t) \\ \eta(x, t) = \int_{B_x} u(z, t) dz \\ y \sim \mathcal{N}(\eta(x, t), \sigma^2) \end{array} \right. \quad (13)$$

where  $u(x, t)$  is in this case the density of horses,  $K(x)$  the habitat-dependent carrying capacity,  $b$  the birth rate and  
 $\mu$  the mortality rate. Population densities are monitored over subdomains  $B_x \subset \Omega$ , each subdomain corresponding to

a horizontal transect with lower-left corner position  $x$ , over which we integrate to obtain observations that are drawn from a Gaussian distribution  $\mathcal{N}(\cdot)$ . We assume all transects have the same width  $l$  (from aerial line of sight) but that the length  $L_x$  will depend on each potential transect. The habitat-dependent carrying capacity was defined as :

$$K(x) = \begin{cases} K_H & \text{if } (x) \text{ in a highly favorable habitat field,} \\ K_O, & \text{otherwise.} \end{cases}$$

223 The population process is parametrized by  $(D, \theta_p)$  with  $\theta_p = (b, \mu, K_H, K_O)$  and the data process depends on  
 224  $\theta_d = (\sigma^2)$ . In this second case study the question we addressed was to find the best  $n$  transects, over  $N$ . We chose to  
 225 simulate a survey of the population performed 7 years after the original monitoring (i.e. 2021). In Beeton & Johnson  
 226 (2019), the mortality rate  $\mu$  was fixed and the carrying capacity  $K(x)$  was estimated independently of the population  
 227 process. Hence, the vector of unknown parameters, for the conservation biology case, was  $\theta = (D, b)$ . One can note  
 228 that  $\sigma$  is classically not part of such optimal design as it can be estimated by residual standard deviation. There was  
 229 no previous estimations for  $\theta$  but there were lower and upper boundaries, expressed as 'low growth and dispersal' and  
 230 'high growth and dispersal' scenarios in the original paper. We used them to set  $\theta_{min}$  and  $\theta_{max}$  used in the on-average  
 231 exact design whereas for local design, we considered  $\theta_{min}$ . Details for all parameters values can be found in appendix  
 232 A.1.

233 In this example, as described in Beeton & Johnson (2019), the experiment consisted in surveying  $N = 32$  aerial  
 234 transects, of different lengths, one time, in the year 2014, to estimate wild horses populations in the australian alps  
 235 (Cairns, 2014).

## 236 5. Criteria for comparing designs and numerical implementation

### 237 5.1. Performance assessment of designs

238 In optimal design, one can quantify the suboptimality of any given designs compared to the D-optimal design, as  
 239 described by eq. 5 and 7, using the notion of the D-efficiency (Ucinski & YangQuan Chen, 2005) which is defined as  
 240 follows :

$$E_D(\xi^n) = \left\{ \frac{\det(M(\theta; \xi^n))}{\det(M(\theta; \xi_*^n))} \right\}^{\frac{1}{q}} \quad (14)$$

241 with  $q$  the number of parameters in  $\theta$ . We can use this measure to compare our designs among themselves, or  
 242 with any other designs, e.g. taken from the literature or even randomly drawn. It can also be of interest to visualize the

243 design properties. Especially, one can review design positions on maps. One might also want to focus on the relationship  
 244 between pair of parameters, using confidence ellipses (Murdoch & Chow, 1996) derived from the inverse of the Fisher  
 245 matrix.

## 246 5.2. Evaluating designs according to a practical constraint

247 When planning an ecological monitoring, the statistical properties of the design is only part of the decision making.  
 248 One important factor for planning the monitoring in the two study cases considered here is the length of travelling  
 249 that induces important costs. In order to integrate this component we consider the tour length of travelling salesman  
 250 solutions (Rosenkrantz, Stearns & Lewis, 1977) where one start at a relevant 'base camp', then tour a design before  
 251 going back to camp. Formally, as defined in Dantzig (1963), we label the traps from  $1, \dots, n$  and define :

$$252 \quad x_{ij} = \begin{cases} 1 & \text{the path goes from trap } i \text{ to trap } j \\ 0 & \text{otherwise} \end{cases}$$

253 Taking  $c_{ij} > 0$  to be the Euclidean distance from trap  $i$  to trap  $j$ , the travelling salesman problem (TSP) can be  
 254 written as the following integer linear programming problem:

$$\begin{aligned} & \min \sum_{i=1}^n \sum_{j \neq i, j=1}^n c_{ij} x_{ij} : & (15) \\ \text{subject to : } & \sum_{i=1, i \neq j}^n x_{ij} = 1 & j = 1, \dots, n; \\ & \sum_{j=1, j \neq i}^n x_{ij} = 1 & i = 1, \dots, n; \\ & \sum_{i \in Q} \sum_{j \neq i, j \in Q} x_{ij} \leq |Q| - 1 & \forall Q \subsetneq \{1, \dots, n\}, |Q| \geq 2 \end{aligned}$$

255 As one can see, the TSP minimize the tour length while ensuring each trap is arrived at from exactly one other  
 256 trap, departed to exactly one other trap and that the solution returned is a single tour and not the union of smaller  
 257 tours. For the second use case, involving lines instead of points, a heuristic is used instead of the problem 15. The tour  
 258 length is roughly approximated by a weighted mean between the diagonal of the bounding box of the transects and  
 259 their cumulated lengths. Finally, the best compromises between statistically efficient designs and travelling lengths  
 260 can be visualized on a Pareto front in the D-efficiency - tour length space (Sheftel, Shoval, Mayo & Alon, 2013). Of  
 261 course, we can note that performance assessment of an experimental design is not limited to the above examples.

### 5.3. Numerical implementation

We implemented different strategies to solve numerically the computationally demanding problems we worked on. As the computation of the on-average design (eq. 11) with a uniform distribution would be cumbersome, we relied on a sampling from a Latin Hypercube, a space-filling design (Pronzato & Müller, 2012), to 'cover' at best a domain with a limited number of sampled points.

Moreover, we noted that one useful feature that come at no cost with our designs is the ability to filter the set of  $N$  possible points (or transects) before searching for a solution. For example, in the agroecological case, we used a regular grid to select  $M$  possible trap locations within fields ( $M \gg N$ ) while excluding road from pitfall trap's possible positions.

In addition, in order to solve the information matrix that requires the derivatives of the deterministic process models we chose to numerically extract them ahead of solving the design, i.e. to do some memoization (Michie, 1968), which proved relevant for the task at hand. For performance, one could also rely on analytic, symbolic or automatic differentiation (adjoint code) if applicable.

Finally, we relied on an excursion algorithm (Fedorov, 1972), with multiple starts, for solving the designs. These algorithms take the particular form of the problem into account. Starting from a non-degenerated design (i.e. with positive determinant), at each iteration  $k$ , exchange one support point by a better one, in the sense of the D-optimality. Considering all  $N * (M - N)$  possible exchanges successively, retaining the best one among these each iteration, it converges to a local solution. Hence, we used multiple starts to get the best local solution as an approximation of a global solution.

Numerical calculations were performed in R (R Core Team, 2019). Using a Runge-Kutta solver combined with the method of lines to solve the reaction-diffusion models (Soetaert, Petzoldt & Setzer, 2010) and some naive parallelization. It took about an hour to solve an on-average design on a Intel® Xeon® E5 with 8 cores 8 Gb of RAM.

## 6. Results

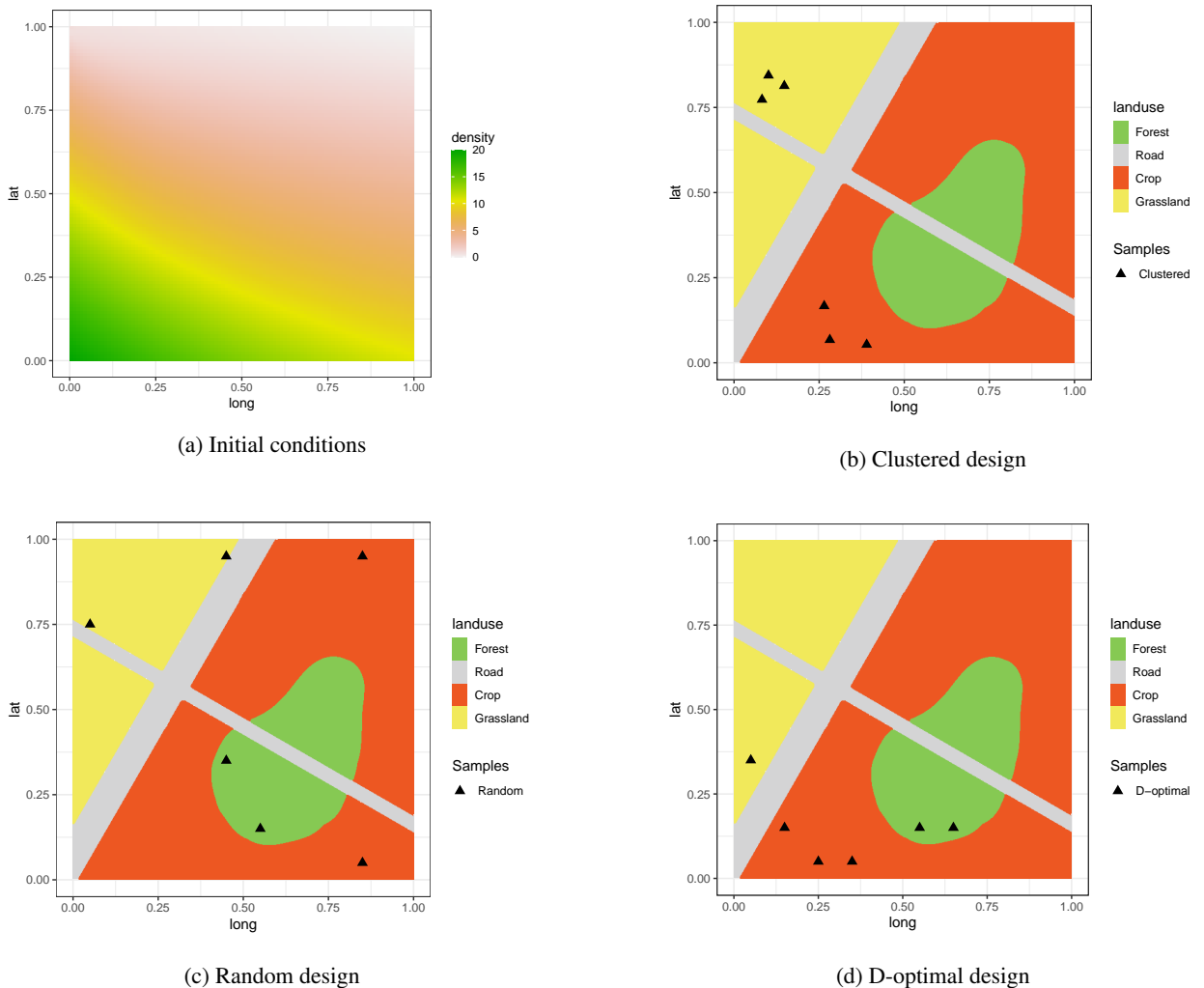
We reduced the size of the original landscapes, and thus the number of sampling points, to reduce the computational cost. For the agroecological case study, we used a virtual agroecosystem, of one km<sup>2</sup>, inspired by Soubeyrand & Roques (2014) with just six pitfall traps sampled over a season. For the conservation biology case study, we used a simplified habitat map, of several thousand km<sup>2</sup>, of a subpart of the australian alps (see supplementary materials of Beeton & Johnson (2019)), sampled one time, 7 years after the first experiment (hence in 2021). For this example, we reduced the number of surveyed transects to eight, chosen among the original thirty two, through D-optimality.

For each use case, we compared empirical designs (e.g. random and/or clustered) to local and on-average D-optimal ones, assuming they both outperform the empirical designs. Theoretically, locally optimal designs should be the most

293 efficient ones regarding the information gain on the processes. However, they require a good knowledge on parameters  
 294 before the survey which is, in practice, rarely the case. Thus, given the usual uncertainty on populations traits and  
 295 considering on-average designs appears more relevant.

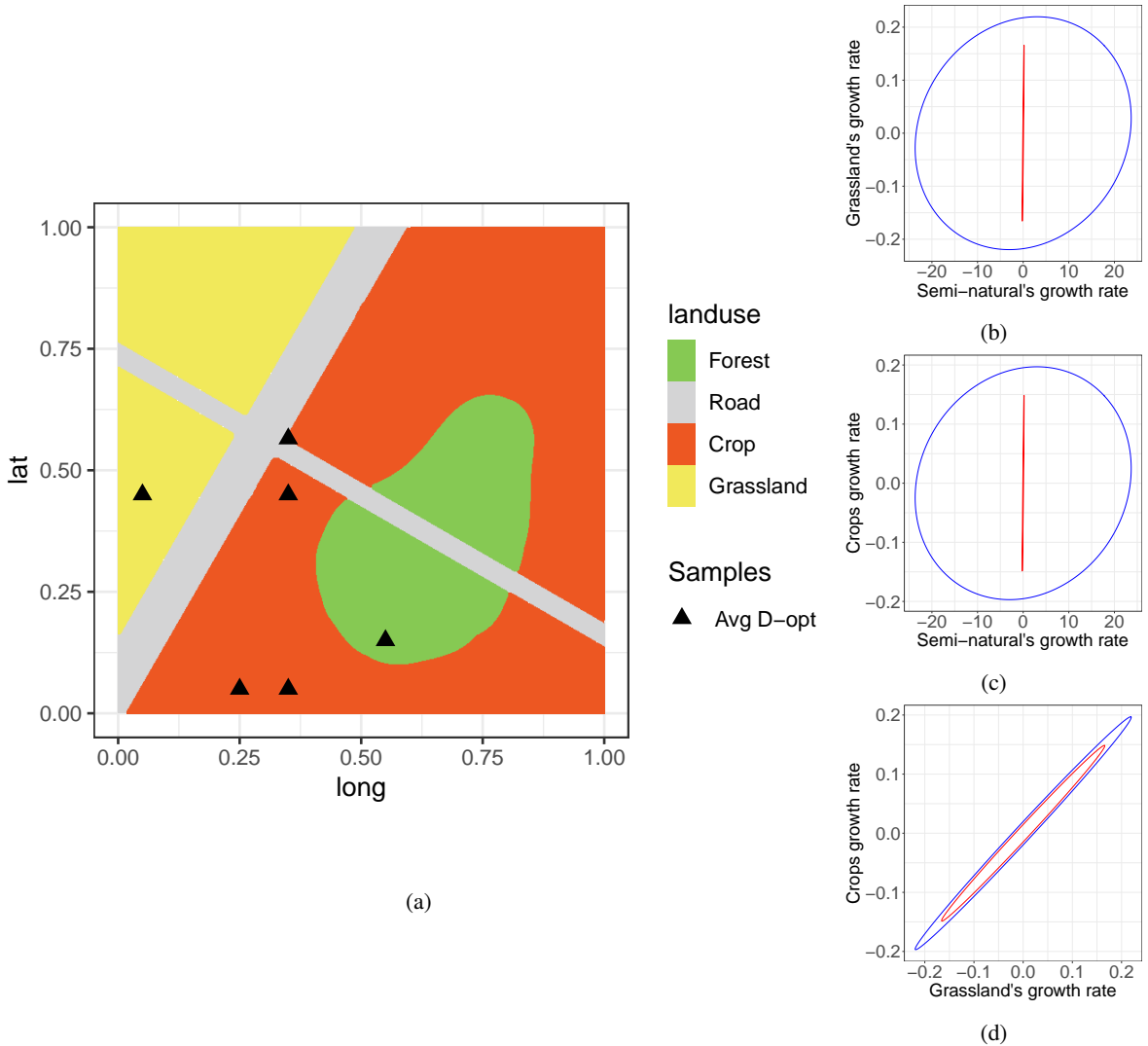
### 296 6.1. Agroecological case study

297 In most studies investigating insect populations in agricultural landscapes, traps are placed according to what we  
 298 call a clustered design. In practice, when a field is chosen, three spatially grouped traps are put in this field as illustrated  
 299 in Fig. 1b. For this case, we compared some properties of local D-optimal, on-average D-optimal, clustered and random  
 300 (i.e. where spatial locations are drawn from a binomial point process) designs. The expected value and the range of  
 301 estimated parameters used for respectively local and on-average designs are given in Appendix A.1.



**Figure 1:** Pitfall trap placements comparisons : (a) initial population densities; (b) clustered design with 3 traps per field (pseudo-replicate) in different fields; (c) randomly positioned pitfall traps ; (d) local D-optimal design

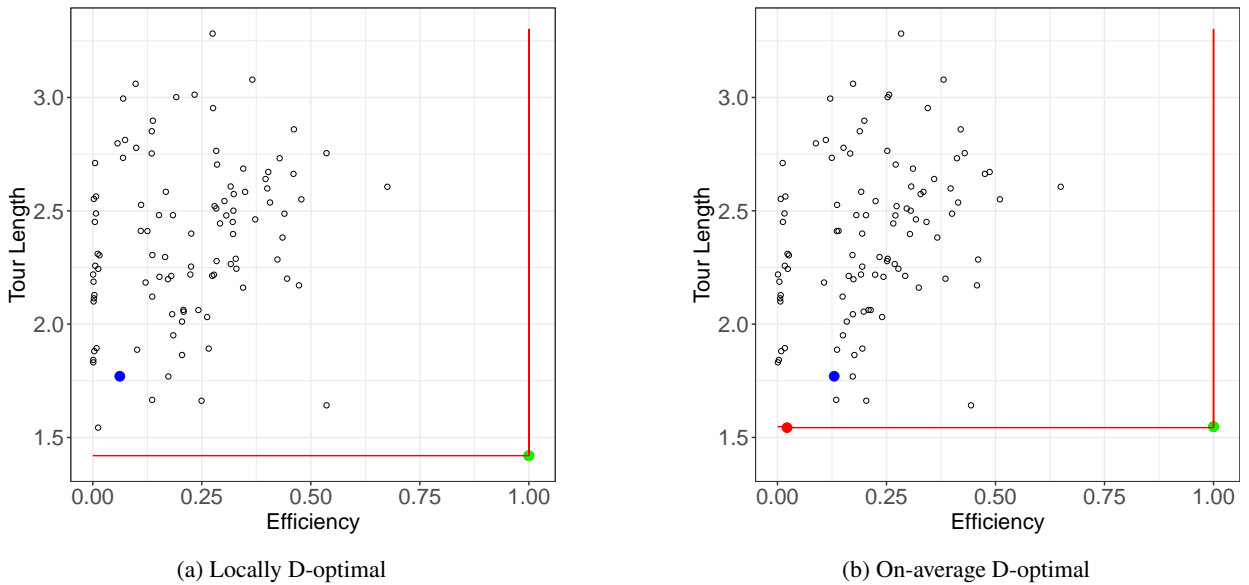
302 As shown in Figure 1, the initial conditions were generated as a gradient from bottom left to top right, and the  
 303 landscape contains 2 crop fields (orange), two grasslands (yellow), one region of semi-natural habitat (e.g. forest patch  
 304 in green) and two crossing roads (gray).  
 305 The local D-optimal design displayed in Figure 1d concentrates the points on the bottom left corner of the landscape  
 306 where the highest slope of initial conditions is. Moreover, it seems to cover all the habitats, even if unequally, except  
 307 roads where the growth rate is set to 0.  
 308



**Figure 2:** On-average exact D-optimal design : (a) position of pitfall traps in case of on-average D-optimal exact design ; (b-d) 95% confidence ellipses for pair of growth rates, estimated by inverting the Fisher Information Matrix for  $\theta$ , for the on-average optimal design (red) and the clustered design (blue) seen in fig. 1.



309 Figure 2 shows the position of the traps in an on-average design and 95% confidence ellipses, estimated by inverting  
 310 the Fisher information matrix, for the pairs of growth rates  $(r_c, r_s)$ ,  $(r_c, r_g)$  and  $(r_s, r_g)$ . Interestingly, compared to the  
 311 local-design, the stencil shape seems less apparent here as we see four points align within the initial condition density  
 312 but two points are farther away and distributed between two different habitats. As expected, the confidence ellipses  
 313 surface are smaller for all pairs of parameters for the on-average D-optimal design compared to the cluster one (Fig.  
 314 2c-d & Appendix A.2 for those related to  $\beta$ ,  $\mu$  and  $D$ ). The uncertainty associated with semi-natural growth rate  $r_s$  is  
 315 severely reduced which is logical given the clustered design did not account for that habitat. In fact, the optimal design  
 316 suggests that only one point within this habitat drastically reduces the variance of its estimates.



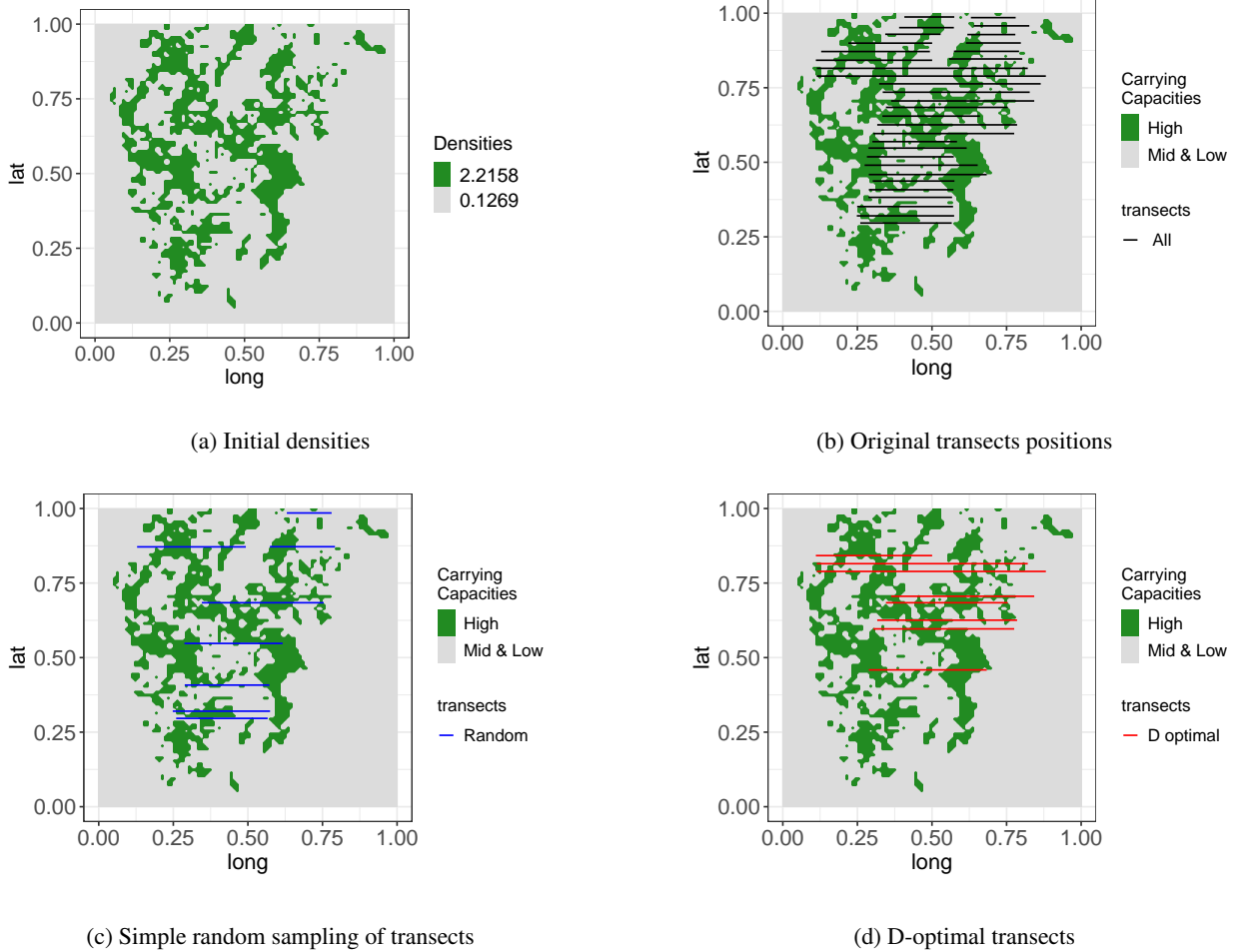
**Figure 3:** Linear approximation of the Pareto front : (a) for the local design; (b) for the on-average design. Each point represents the D-efficiency (x) and tour length (y) of a design. Black circles represent 100 random samplings (the same set is projected on each panel), blue dots represents the clustered sampling, green dots are optimal and red lines and dots form the linear approximation to the pareto fronts. All the designs D-efficiencies are evaluated locally (a) or on-average (b).

317 The comparison of the efficiency of the designs in relation to the tour length is presented in figure 3 for both local  
 318 and on-average designs. Random designs are represented by black circles (there are a hundred of them), the clustered  
 319 design by a blue dot, the D-optimal design by a green dot and if at least one red dot exist, it is a random sampling  
 320 that is pareto optimal for at least one dimension. One can see that the clustered design is, of course, less statistically  
 321 efficient than the D-optimal but also that it has a longer tour. In fact, the D-optimal designs here are dominant both in  
 322 statistical and tour length. Several simulations of random designs show that it is not always the case but the optimal  
 323 designs are still among the shortest (data not shown). One can also note that most random designs are more D-efficient  
 324 than the clustered design. This is probably because they either place traps within semi-natural habitat and/or simply  
 325 cover more ground, with a longer tour length. As in this study we drew only one clustered empirical design, the solid

326 comparison between random and clustered sampling strategies, which was not the scope of this work, would require  
 327 more simulations.

## 328 6.2. Conservation biology case study

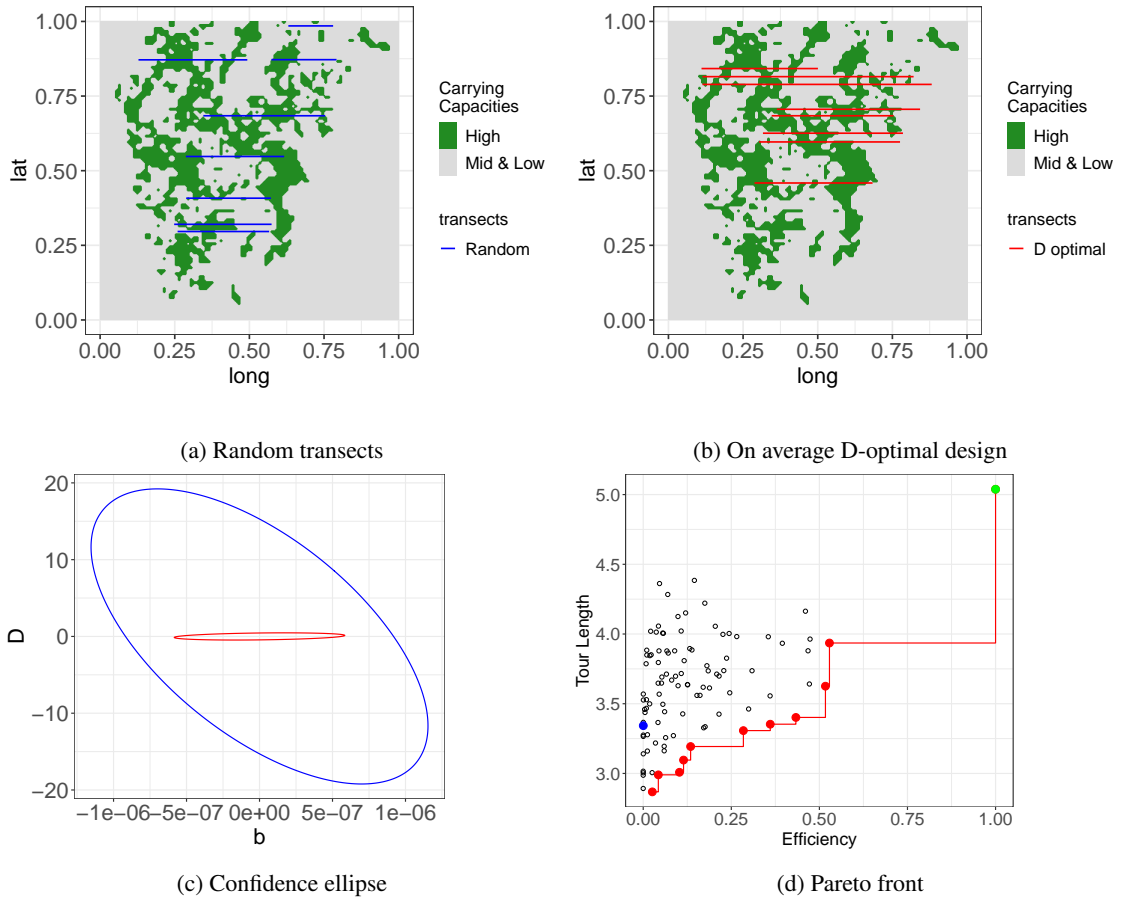
329 In this example the design problem consists in searching  $n = 8$  aerial transects, of different lengths, chosen among  
 330  $N=32$  reported in Cairns (2014). The expected values and the range of parameters  $\theta$  are given in Appendix A.1.



**Figure 4:** Transects comparisons: In the original study, high initial densities were linked with a high carrying capacity while mid and low densities were linked with another carrying capacity (see also Appendix A.1). We can see here (a) the initial population densities (b) then all 32 original transects (in black) over the carrying capacities ; (c) some random subset of 8 of the transects (in blue) ; (d) finally a locally D-optimal design of 8 transects for  $\theta_{min}$  (i.e. 'low growth and dispersal' scenario) (in red)

331 As shown in figure 4a the monitored area has less high quality habitats (green) than average and low quality ones  
 332 (grey). The original transect coverage was extremely dense (Fig. 4b) and it appears natural to reduce it, for instance by  
 333 taking only  $\frac{1}{4}$  of the possible transects. As short transects are frequent in the original empirical distribution of transects,

334 there are a few chances of randomly choosing the longer ones (Fig. 4c). When defining a local D-optimal design with  
 335 a 'low growth and dispersal' scenario, the longer and more informative transects appear to be favored (Fig. 4d).



**Figure 5:** On-average exact D-optimal design : (a) random transects; (b) on-average D-optimal ; (c) 95% confidence ellipse of  $b$  and  $D$ , blue is for the random design figured in (a) and red for the optimal design, both evaluated at  $\theta_{min}$  and, finally, (d) a linear approximation of the Pareto front with D-efficiency evaluated at  $\theta_{min}$  (color coding and shapes are the same as in figure 3)

336 The on-average D-optimal design (5b) appears to be similar to the local design (Fig. 4d). It seems to favor longer  
 337 transects and, in that case, only one transect on the southern part of the map. The spatial extent of the optimal transects  
 338 is smaller than most random transects but, as it favors longer transects, the tour length, forcibly passing by transects, is  
 339 actually longer than most (Fig. 5d). As expected, D-optimality reduces the confidence ellipse of estimated parameters  
 340 compared to a random design (Fig. 5c). The variance reduction appears to be greater for the diffusion coefficient  $D$   
 341 than for the growth rate  $b$ . Finally, the Pareto front displayed in Figure 5d to assess the compromise between design  
 342 efficiency and tour length, shows that the on-average D-optimal transects dominate the D-efficiency but is here the  
 343 worst choice in term of tour length. This means that there are multiple point on the Pareto front that could be candidate

344 if a compromise has to be chosen between D-efficiency and tour length.

## 346 7. Discussion

347 In this study we addressed the question of where to make observations within a large spatial domain to efficiently  
348 capture the changes in space and time of a population described by a reaction-diffusion model. By considering two  
349 example systems, i.e. the dynamics of carabid beetles in agricultural landscapes and wild horses in the Australian  
350 Alps, we showed that the use of the optimal design of experiments framework can be useful to find optimal locations  
351 for trapping or counting individuals that maximize the information on the population for parameter estimation. In  
352 particular, we were able to produce spatial designs that outperform both those used in the initial studies and random  
353 designs. More precisely we first obtained local D-optimal designs that assume that the expected values of parameters are  
354 known before planning data collection. Then, as this situation is generally unlikely to occur in the population ecology,  
355 we investigated the situations when only bounds around population parameters are known considering on-average  
356 D-optimal designs. While local designs appeared to be obviously the most efficient for minimizing the variance of  
357 parameters estimates, our results points out the usefulness of optimal design without strong *a priori* to think spatial  
358 monitoring so it will enable efficient model fitting.

359 Although the question of optimized survey in ecology or epidemiology is still seldom considered, it has already  
360 been addressed in several studies using various frameworks. For instance, Williams et al. (2018) proposed a framework  
361 that produces designs that minimizes prediction uncertainty, Cook et al. (2008) calculated optimal observation times  
362 for botanical epidemiology experiments by maximizing the Kullback-Leibler divergence from the posterior to the  
363 prior, or Bourhis, Bell, van den Bosch & Milne (2021) who proposed a strategy for choosing insect traps locations by  
364 minimizing uncertainty in neural networks outputs. Here, we considered the optimal design of experiments framework  
365 that is briefly introduced in part 3. While its use is common for solving experimental design issues in environmental  
366 sciences or systems biology (Steiert, Raue, Timmer & Kreutz, 2012), it is poorly considered in ecology. The monitoring  
367 of population on large spatial domain is actually similar to the problem of optimal sensors placement for spatially  
368 distributed systems (Ucinski, 2004).

369 Despite the recent development of devices for increasing ecological monitoring, in practice the survey of population  
370 in large spatial domains remains challenging and a major bottleneck to the study of how populations change in space  
371 and time and the design of efficient management strategies (Nichols & Williams, 2006; Lindenmayer & Likens, 2010).  
372 As stated above, statistical frameworks developed to optimize how and where one should make observations can be  
373 used to design more efficient population surveys. This model-based design of experiments can also take into account  
374 several practical constraints. In our study we chose to consider the tour length between traps for collecting carabid

375 bettles and aerial transects from which wild horses are enumerated. Then we used a Pareto front to asses how designs  
376 efficiency changes with the tour length. As illustrated in Fig. 5d, it is possible to find good candidate that maximizes  
377 design efficiency while minimizing tour length so one can choose those that fit the compromises that are acceptable. For  
378 the agroecology example, we found that the best design was also minimizing the tour length. This may be explained  
379 by the smaller spatial extent of the domain compared to the wild horse case. In fact, it is likely that increasing the  
380 spatial domain would create a more progressive Pareto front between design efficiency and tour length. Depending  
381 on the considered problem, other constraints could be included in the approach to find designs that are statistically  
382 efficient and minimized the constraints. For instance one could imagine several travellers and seek for the strategy that  
383 minimizes the cost of the survey, considering both human and transport costs. While, in our opinion, finding monitoring  
384 designs that provide maximum information on population spread is an important question on which the community of  
385 modellers in ecology should concentrate more, integrating realistic constraints in model-based design is likely to catch  
386 the interest of field ecologists and practitioners. Moreover, in our two study cases we focused on the spatial aspect of  
387 monitoring assuming that the only degree of freedom is the survey designs was the locations of observations whereas  
388 the time of observations was fixed. This assumption could be relaxed and both temporal and spatio-temporal D-optimal  
389 designs may be obtained in further studies. Optimal times at which observations should be done may be obtained by  
390 solving local or on-average D-optimality problems before the survey starts. In addition, if applicable, the problem can  
391 also be tackled sequentially as an alternative to the on-average design. In this case, optimal groups of observations are  
392 proposed, over time, after having performed a parameter estimation on already acquired data, this approach is know  
393 as a sequential design (Chernoff, 1959).

394 Apart the theoretical background, the main obstacle to a widespread usage of optimal design of experiments in  
395 spatially explicit systems is the computational cost. Using usual algorithms and computing strategies (i.e. memoization  
396 and parallelization), in this study we were able to solve local and an-average D-optimality problems on reaction-  
397 diffusion models with a reasonable amount of time. However, increasing the complexity of the problem, for instance  
398 by adding more practical or economic constraints or the temporal dimension of sampling, may require more advanced  
399 computing methods and grid computing. Structural and practical identifiability issues may also arise and make the  
400 problem even more complex. Optimal design is useful for maximizing the information gain from data collection,  
401 however the modeller has to ensure that enough information is available so the parameters can be determined. In this  
402 study, before seeking for an optimal design, we assessed the practical identifiability of the model by multiple start of  
403 a numerical local approach (Walter & Pronzato, 1997) within each relevant domain  $\Theta$ .

404 The issue of initial conditions is well known by mathematicians and modellers working with dynamic models. In  
405 some conditions, it is possible to estimate initial conditions from monitoring data. Nevertheless, in most studies one  
406 has to make strong hypothesis and find a way to fix it before performing a statistical inference, or seeking an optimal

407 design. Regarding spatial ecology, the most frequent approach is perhaps to use past (or early data) with a regression  
 408 model to somehow estimate the distribution of population density at the beginning of the study. In our study, we  
 409 did not include the initial conditions in the designs problems. It is very likely that optimal designs are dependent  
 410 on the initial conditions. This may be assessed by augmenting the problem and exploring how changes in the initial  
 411 conditions impact optimal-designs. Yet, augmenting the problem for taking care of initial conditions has its limits.  
 412 To cite in extenso Uciński (2018), it is "[a] very prospective direction [where] the infinite dimensional nature of the  
 413 resulting parameter space is inherently associated with the ill-posedness which means that even low noise in the data  
 414 may make the estimates extremely unstable.". The reader could refer to Alexanderian, Petra, Stadler & Ghattas (2014)  
 415 and Alexanderian, Petra, Stadler & Ghattas (2016) for further information.

416 In this article we focus on the  $D$  criterion but there exist different criteria (e.g. different functions  $\phi(\cdot)$ ) that can  
 417 be applied to the information matrix. For example, one can seek to minimize the average variance of the estimates  
 418 by minimizing the negative of the trace of the inverse of the Fisher information matrix (i.e.  $A$ -optimality). Here, we  
 419 focused on the  $D$  criterion because (i) through the equivalence theorem (Kiefer, 1974), we know that main classical  
 420 criteria are linked in some ways (e.g.  $A$  and  $D$ ), (ii) the  $D$  criterion is invariant by reparameterization (e.g. helpful to  
 421 control parameters positivity) and (iii) there exist variants of the  $D$  criterion that are useful for a variety of purposes.  
 422 For instance, if one is interested in model selection, the  $D_s$  criterion enable to focus the problem on the parameters of  
 423 interest versus 'nuisance' parameters, hence being a canonical criterion in nested model selection (Atkinson & Cox,  
 424 1974). For more general purpose model selection in case of gaussian observations, the generalized  $DT$  criterion is a  
 425 criterion that can be described as a normalized arithmetic mean of  $D$  criterions and  $L_2$  distance between predictions,  
 426 handling both optimal parameters estimation and model selection (Atkinson, 2008).

427 One can also note that, quite often, the aim of ecological surveys is not only to estimate reaction-diffusion model  
 428 parameters but rather to monitor the state of the population and provide abundances. For such goal, it would be better to  
 429 use a criterion more directly related to predictions. For instance, the  $G$  criterion seeks to minimize the maximum entry  
 430 in the diagonal of the hat matrix, minimizing the maximum variance of the predicted values i.e.  $u(\hat{x}, t)$ . As Pázman &  
 431 Pronzato (2014) refined the  $G$  criterion for nonlinear homoscedastic or heteroscedastic regression models, we think an  
 432 on-average  $G$ -optimal design would be more suitable than  $D$ -optimal designs to focus on population state monitoring.  
 433 Even if knowing model's parameters and knowing  $u$  is related, it corresponds to different mathematical objectives (e.g.  
 434  $G$  versus  $D$  optimal designs) that may produce distinct designs. One can note that some limitations might have to be  
 435 worked through for the  $G$ -optimal design in case of the generalized regression model (i.e. with any kind of data model)  
 436 and the numerical aspects of the problem.

437 To finish with, the purpose of this article was to push the discussion on the statistical efficiency of experimental  
 438 designs linked with spatial ecological models, as question that is yet seldom expressed this way. The main goal of a

439 field ecologist is to sample 'as much as possible' within the constraints of its budgets (monetary and manpower), but  
440 this strategy can either fail to capture the essential information on the population or be very expensive. In addition to  
441 the numerous studies that had demonstrated that formalizing ecological processes into mathematical models and using  
442 them to analyze empirical data offers a mean for improving our understanding of populations spread, we hope this work  
443 points out that the difference between optimal and non-optimal monitoring can be significant in term of information  
444 gained, and thus that model-based design of ecological survey is a promising path that could also contribute to reduce  
445 environmental costs of samples collection, storage and processing.

#### 446 **CRedit authorship contribution statement**

447 **Nicolas Parisey:** Implemented the framework, ran the numerical experiments and provided focus on optimal  
448 design. **Melen Leclerc:** provided insight on ecological modeling. **Katarzyna Adamczyk-Chauvat:** provided fruitful  
449 guidance on statistics. **All authors** designed the numerical experiments and the overall framework, drafted, managed,  
450 reviewed and approved the manuscript.

#### 451 **Software and data availability**

452 All our results can be reproduced, extended and adapted to other use cases by cloning the code and data available  
453 at <https://github.com/nparisey-pro/LANDoE> under GPL license. The software was developed using R (R Core  
454 Team, 2019) with a dozen package dependencies. Packages list, versions and tested operating systems are listed at the  
455 git repository.

#### 456 **Competing interests.**

457 The authors declare that they have no known competing financial interests or personal relationships that could have  
458 appeared to influence the work reported in this paper.

#### 459 **Funding.**

460 The authors thank the ANR project "Clonix 2D" (ANR-18-CE32-0001), and its leader Solenn Stoeckel, for its  
461 financial support of this work.

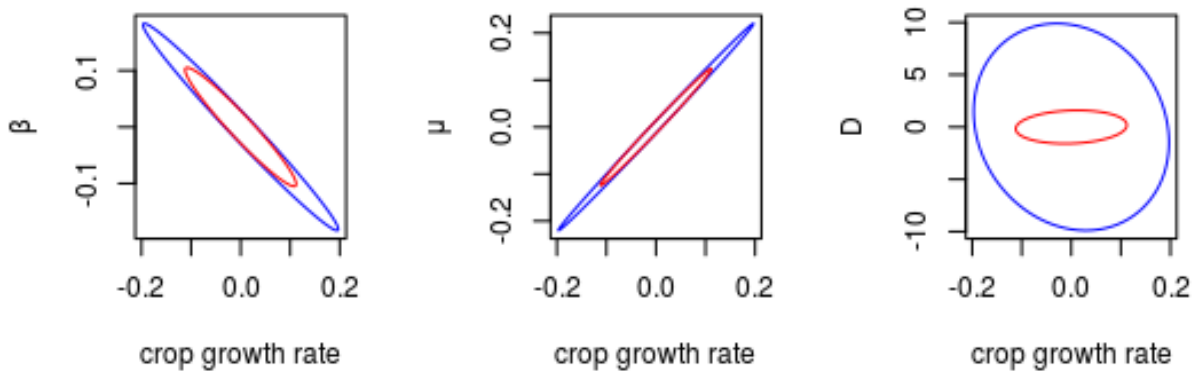
#### 462 **Acknowledgements.**

463 The authors thank Jean-Sébastien Pierre, Jacques Baudry and Solenn Stoeckel for useful discussions about this  
464 work. The authors want to thank especially Professor Dariusz Uciński for a brief but very supportive discussion about  
465 an early version of this work.

466 **A. Appendix**467 **A.1. Use case's parameters****Table 1**

Parameters used for the use cases when searching for local and on-average optimal designs. The symbol '–' means 'not applicable' e.g. when used for explored min or max, it means those parameters were not part of the optimal design problem so they were kept at their original values. Reference [1] is (Parisey et al., 2016) and reference [2] is (Beeton & Johnson, 2019).

Parameters	Original values	Explored min	Explored max	units	source
$D$	76.1	19.025	304.4	$\text{m}^2 \cdot \text{day}^{-1}$	[1]
$\beta$	0.123	0.111	0.135	–	[1]
$\mu$	0.210	0.115	0.42	$\text{day}^{-1}$	[1]
$r_s$	0.155	0.144	0.166	$\text{day}^{-1}$	[1]
$r_g$	0.304	0.293	0.315	$\text{day}^{-1}$	[1]
$r_c$	0.385	0.374	0.396	$\text{day}^{-1}$	[1]
$r_r$	0	–	–	$\text{day}^{-1}$	[1]
$\eta$	$10^{-4}$	–	–	–	[1]
$\tau$	5	–	–	day	[1]
$K_H$	2.462	–	–	horse	[2]
$K_O$	0.141	–	–	horse	[2]
$D$	–	2	30	nondimensional	[2]
$b$	–	0.16	0.27	$\text{year}^{-1}$	[2]
$\sigma$	$10^{-3}$	–	–	$\text{horse}^2$	arbitrarily low

468 **A.2. Supplementary confidence ellipses**

**Figure 5:** Confidence ellipses at 95% for the agroecological case study, where parameters  $\beta$ ,  $\mu$  and  $D$  are paired with the crop growth rate.



## References

- 469
- 470 Abboud, C., Bonnefon, O., Parent, E., & Soubeyrand, S. (2019). Dating and localizing an invasion from post-introduction data and a coupled  
471 reaction–diffusion–absorption model. *Journal of mathematical biology*, *79*, 765–789.
- 472 Alexanderian, A., Petra, N., Stadler, G., & Ghattas, O. (2014). A-Optimal Design of Experiments for Infinite-Dimensional Bayesian Linear Inverse  
473 Problems with Regularized  $\ell_0$ -Sparsification. *SIAM Journal on Scientific Computing*, *36*, A2122–A2148. URL: [http://epubs.siam.org/](http://epubs.siam.org/doi/10.1137/130933381)  
474 [doi/10.1137/130933381](http://epubs.siam.org/doi/10.1137/130933381). doi:10.1137/130933381.
- 475 Alexanderian, A., Petra, N., Stadler, G., & Ghattas, O. (2016). A fast and scalable method for A-optimal design of experiments for infinite-  
476 dimensional Bayesian nonlinear inverse problems. *SIAM Journal on Scientific Computing*, *38*, A243–A272.
- 477 Atkinson, A. C. (2008). DT-optimum designs for model discrimination and parameter estimation. *Journal of Statistical Planning and Inference*, *138*,  
478 56–64. URL: <http://www.sciencedirect.com/science/article/pii/S0378375807001954>. doi:10.1016/j.jspi.2007.05.024.
- 479 Atkinson, A. C., & Cox, D. R. (1974). Planning Experiments for Discriminating Between Models. *Journal of the Royal Statistical Society. Series*  
480 *B (Methodological)*, *36*, 321–348. URL: <https://www.jstor.org/stable/2984923>.
- 481 Atkinson, A. C., Donev, A., & Tobias, R. (2007). *Optimum Experimental Designs, With SAS*. OUP Oxford.
- 482 Atkinson, A. C., Fedorov, V. V., Herzberg, A. M., & Zhang, R. (2014). Elemental information matrices and optimal experimental design  
483 for generalized regression models. *Journal of Statistical Planning and Inference*, *144*, 81–91. URL: [http://www.sciencedirect.com/](http://www.sciencedirect.com/science/article/pii/S0378375812003060)  
484 [science/article/pii/S0378375812003060](http://www.sciencedirect.com/science/article/pii/S0378375812003060). doi:10.1016/j.jspi.2012.09.012.
- 485 Beeton, N. J., & Johnson, C. N. (2019). Modelling horse management in the Australian Alps. *Ecological Management & Restoration*, *20*, 57–62.  
486 URL: <https://onlinelibrary.wiley.com/doi/abs/10.1111/emr.12350>. doi:10.1111/emr.12350.
- 487 Bellot, B., Poggi, S., Baudry, J., Bourhis, Y., & Parisey, N. (2018). Inferring ecological processes from population signatures: A simulation-based  
488 heuristic for the selection of sampling strategies. *Ecological Modelling*, *385*, 12–25.
- 489 Bourhis, Y., Bell, J. R., van den Bosch, F., & Milne, A. E. (2021). Artificial neural networks for monitoring network optimisation—a practical  
490 example using a national insect survey. *Environmental Modelling & Software*, *135*, 104925.
- 491 Cairns (2014). *Feral Horses in the Australian Alps: the Design and Analysis of Surveys Conducted in April-May, 2014*. Technical Report G.E. &  
492 S.C. Cairns Consulting Pty. Ltd.
- 493 Chernoff, H. (1959). Sequential design of experiments. *The Annals of Mathematical Statistics*, *30*, 755–770. URL: [http://projecteuclid.](http://projecteuclid.org/euclid.aoms/1177706205)  
494 [org/euclid.aoms/1177706205](http://projecteuclid.org/euclid.aoms/1177706205). doi:10.1214/aoms/1177706205.
- 495 Clark, J. S., & Bjørnstad, O. N. (2004). Population time series: process variability, observation errors, missing values, lags, and hidden states.  
496 *Ecology*, *85*, 3140–3150.
- 497 Cliff, O. M., Saunders, D. L., & Fitch, R. (2018). Robotic ecology: Tracking small dynamic animals with an autonomous aerial vehicle. *Science*  
498 *Robotics*, *3*.
- 499 Cook, A. R., Gibson, G. J., & Gilligan, C. A. (2008). Optimal observation times in experimental epidemic processes. *Biometrics*, *64*, 860–868.
- 500 Dantzig, G. B. (1963). *Linear Programming and Extensions*.
- 501 Fedorov, V. V. (1972). *Theory Of Optimal Experiments*. Elsevier.
- 502 Fisher, R. (1935). *The design of experiments. 1935*. Edinburgh: Oliver and Boyd.
- 503 Fornberg, B., & Sloan, D. M. (1994). A review of pseudospectral methods for solving partial differential equa-  
504 tions. *Acta Numerica*, *3*, 203–267. URL: [https://www.cambridge.org/core/journals/acta-numerica/](https://www.cambridge.org/core/journals/acta-numerica/article/a-review-of-pseudospectral-methods-for-solving-partial-differential-equations/E921CA187B5F35E0BEF9ACAA285C083B)  
505 [article/a-review-of-pseudospectral-methods-for-solving-partial-differential-equations/](https://www.cambridge.org/core/journals/acta-numerica/article/a-review-of-pseudospectral-methods-for-solving-partial-differential-equations/E921CA187B5F35E0BEF9ACAA285C083B)  
506 [E921CA187B5F35E0BEF9ACAA285C083B](https://www.cambridge.org/core/journals/acta-numerica/article/a-review-of-pseudospectral-methods-for-solving-partial-differential-equations/E921CA187B5F35E0BEF9ACAA285C083B). doi:10.1017/S0962492900002440.

- 507 Hooten, M. B., Wikle, C. K., Sheriff, S. L., & Rushin, J. W. (2009). Optimal spatio-temporal hybrid sampling designs for ecological monitoring.  
508 *Journal of Vegetation Science*, 20, 639–649. URL: [http://onlinelibrary.wiley.com/doi/10.1111/j.1654-1103.2009.01040.x/](http://onlinelibrary.wiley.com/doi/10.1111/j.1654-1103.2009.01040.x/full)  
509 full.
- 510 Hundsdorfer, W., & Verwer, J. (2003). *Numerical Solution of Time-Dependent Advection-Diffusion-Reaction Equations* volume 33 of *Springer*  
511 *Series in Computational Mathematics*. Berlin, Heidelberg: Springer Berlin Heidelberg. URL: [http://link.springer.com/10.1007/](http://link.springer.com/10.1007/978-3-662-09017-6)  
512 978-3-662-09017-6.
- 513 Kiefer, J. (1974). General equivalence theory for optimum designs (approximate theory). *The Annals of Statistics*, 2, 849–879. URL: <http://www.jstor.org/stable/2958055>.
- 514 //www.jstor.org/stable/2958055.
- 515 Lehmann, E. L., & Casella, G. (1998). *Theory of Point Estimation*. Springer Texts in Statistics (2nd ed.). New York: Springer-Verlag. URL:  
516 <https://www.springer.com/br/book/9780387985022>.
- 517 Lindenmayer, D. B., & Likens, G. E. (2010). The science and application of ecological monitoring. *Biological conservation*, 143, 1317–1328.
- 518 Louvrier, J., Papaix, J., Duchamp, C., & Gimenez, O. (2020). A mechanistic–statistical species distribution model to explain and forecast wolf  
519 (*canis lupus*) colonization in south-eastern france. *Spatial Statistics*, 36, 100428.
- 520 Malebranche, H. (1988). Simultaneous state and parameter estimation and location of sensors for distributed systems, . 19, 1387–1405. URL:  
521 <https://doi.org/10.1080/00207728808964045>. doi:10.1080/00207728808964045.
- 522 Marrec, R., Badenhausser, I., Bretagnolle, V., Börger, L., Roncoroni, M., Guillon, N., & Gauffre, B. (2015). Crop succession and habitat preferences  
523 drive the distribution and abundance of carabid beetles in an agricultural landscape. *Agriculture, Ecosystems & Environment*, 199, 282–289.  
524 URL: <http://www.sciencedirect.com/science/article/pii/S016788091400468X>. doi:10.1016/j.agee.2014.10.005.
- 525 Michie, D. (1968). "Memo" Functions and Machine Learning. *Nature*, 218, 19. URL: <https://www.nature.com/articles/218019a0>.  
526 doi:10.1038/218019a0.
- 527 Murdoch, D. J., & Chow, E. D. (1996). A graphical display of large correlation matrices. *The American Statistician*, 50, 178–180. doi:10.1080/  
528 00031305.1996.10474371.
- 529 Nichols, J. D., & Williams, B. K. (2006). Monitoring for conservation. *Trends in ecology & evolution*, 21, 668–673.
- 530 Parisey, N., Bourhis, Y., Roques, L., Soubeyrand, S., Ricci, B., & Poggi, S. (2016). Rearranging agricultural landscapes towards habitat quality  
531 optimisation: In silico application to pest regulation. *Ecological Complexity*, . URL: [http://dx.doi.org/10.1016/j.ecocom.2016.07.](http://dx.doi.org/10.1016/j.ecocom.2016.07.003)  
532 003. doi:10.1016/j.ecocom.2016.07.003.
- 533 Pronzato, L., & Müller, W. G. (2012). Design of computer experiments: space filling and beyond. *Statistics and Computing*, 22, 681–701. URL:  
534 <https://doi.org/10.1007/s11222-011-9242-3>. doi:10.1007/s11222-011-9242-3.
- 535 Pázman, A., & Pronzato, L. (2014). Optimum design accounting for the global nonlinear behavior of the model. *The Annals of Statistics*, 42,  
536 1426–1451. doi:10.1214/14-AOS1232.
- 537 R Core Team (2019). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing Vienna, Austria. URL:  
538 <http://www.R-project.org/>.
- 539 Roques, L., & Bonnefon, O. (2016). Modelling population dynamics in realistic landscapes with linear elements: a mechanistic-statistical reaction-  
540 diffusion approach. *PLoS one*, 11.
- 541 Roques, L., Soubeyrand, S., & Rousselet, J. (2011). A statistical-reaction–diffusion approach for analyzing expansion processes. *Journal of*  
542 *Theoretical Biology*, 274, 43–51.
- 543 Rosenkrantz, D., Stearns, R., & Lewis, I., P. (1977). An Analysis of Several Heuristics for the Traveling Salesman Problem. *SIAM Journal on*  
544 *Computing*, 6, 563–581. URL: <https://epubs.siam.org/doi/10.1137/0206041>. doi:10.1137/0206041.

- 545 Sheftel, H., Shoval, O., Mayo, A., & Alon, U. (2013). The geometry of the pareto front in biological phenotype space. *Ecol-*  
 546 *ogy and Evolution*, 3, 1471–1483. URL: <https://onlinelibrary.wiley.com/doi/abs/10.1002/ece3.528>. doi:10.1002/ece3.528.  
 547 arXiv:<https://onlinelibrary.wiley.com/doi/pdf/10.1002/ece3.528>.
- 548 Silvey, S. (2013). *Optimal design: an introduction to the theory for parameter estimation* volume 1. Springer Science & Business Media.
- 549 Soetaert, K., Petzoldt, T., & Setzer, R. (2010). Solving differential equations in r: Package desolve. *Journal of Statistical Software, Articles*, 33,  
 550 1–25. URL: <https://www.jstatsoft.org/v033/i09>. doi:10.18637/jss.v033.i09.
- 551 Soubeyrand, S., & Roques, L. (2014). Parameter estimation for reaction-diffusion models of biological invasions. *Population Ecology*, 56, 427–434.  
 552 URL: <http://link.springer.com/article/10.1007/s10144-013-0415-0>. doi:10.1007/s10144-013-0415-0.
- 553 Steiert, B., Raue, A., Timmer, J., & Kreutz, C. (2012). Experimental design for parameter estimation of gene regulatory networks. *PLOS ONE*,  
 554 7, e40052. URL: <https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0040052>. doi:10.1371/journal.  
 555 pone.0040052.
- 556 Sun, N.-Z. (1999). *Inverse Problems in Groundwater Modeling*. Theory and Applications of Transport in Porous Media. Springer Netherlands.  
 557 URL: <https://www.springer.com/gp/book/9780792329879>. doi:10.1007/978-94-017-1970-4.
- 558 Ucinski, D. (2004). *Optimal measurement methods for distributed parameter system identification*. CRC press.
- 559 Uciński, D. (2018). Optimum experimental design for infinite dimensional inverse problems. In *Proc. of Workshop on Design of Experiments :*  
 560 *New Challenges*.
- 561 Ucinski, D., & YangQuan Chen (2005). Time-Optimal Path Planning of Moving Sensors for Parameter Estimation of Distributed Systems. In  
 562 *Proceedings of the 44th IEEE Conference on Decision and Control* (pp. 5257–5262). doi:10.1109/CDC.2005.1582997.
- 563 Walter, E., & Pronzato, L. (1997). *Identification of parametric models from experimental data*. Springer.
- 564 Weinstein, B. G. (2018). A computer vision for animal ecology. *Journal of Animal Ecology*, 87, 533–545.
- 565 Wikle, C. K. (2003). Hierarchical models in environmental science. *International Statistical Review*, 71, 181–199.
- 566 Williams, P. J., Hooten, M. B., Womble, G. G., Jamie N. and Esslinger, & Bower, M. R. (2018). Monitoring dynamic spatio-temporal ecological  
 567 processes optimally. *Ecology*, 99, 524–535. URL: <https://esajournals.onlinelibrary.wiley.com/doi/abs/10.1002/ecy.2120>.  
 568 doi:10.1002/ecy.2120.