1 **This is the pre-peer reviewed version of the following article:** Louvet, A.,
2 Machon, N., Mihoub, J. B., Robert, A. (2021). Detecting seed bank influence on plant
3 metapopulation dynamics. *Methods in Ecology and Evolution*, 12(4), 655-664., **which**
4 **has been published in final form at** <https://doi.org/10.1111/2041-210X.13547>.
5 **This article may be used for non-commercial purposes in accordance with**
6 **Wiley Terms and Conditions for Use of Self-Archived Versions.**

Detecting seed bank influence on plant metapopulation dynamics

Apolline Louvet¹, Nathalie Machon², Jean-Baptiste Mihoub², Alexandre Robert²

¹ Centre de Mathématiques Appliquées (CMAP), Ecole Polytechnique, Centre National de la Recherche Scientifique, Institut Polytechnique de Paris, Route de Saclay, 91128 Palaiseau Cedex, France

² Centre d'Ecologie et des Sciences de la Conservation (CESCO), Muséum National d'Histoire Naturelle, Centre National de la Recherche Scientifique, Sorbonne Université, Paris, France

Corresponding author : Apolline Louvet (apolline.louvet@polytechnique.edu)

Running headline : Seed bank detection in plant metapopulations

Abstract

1. Seed banks are known to play a key role in plant metapopulations. However, detecting seed banks remains challenging and requires intense monitoring efforts. Assessing the genuine effect of seed banks on plant metapopulation dynamics may offer a much easier while still biologically relevant way to overcome this issue.

2. In this study, we developed a new metric : the Seed Bank Characteristic Event (SBCE) probability. Instead of detecting seed bank directly, the SBCE probability measures seed bank contribution to the observed metapopulation dynamics. Exploring seed bank parameters (colonization, germination and seed bank death probabilities, initial proportion of patches containing a seed bank), a wide range of monitoring durations (from 3 to 10 years) and number of patches in the metapopulation (from 10 to 1000 patches), we examined the conditions under which the SBCE probability is correctly estimated. To test the robustness of our approach, we further introduced false negatives, false positives or parameter heterogeneity between patches. Finally, we applied the SBCE probability method to the monitoring of tree bases plant species in Paris, France, to assess the applicability of the method to real-world datasets. We studied the influence of species traits and environmental characteristics, in order to increase understanding of plant metapopulation dynamics within an urban environment.

3. Our results indicate that the SBCE probability is well estimated when enough monitoring years or number of patches are considered, and for rates of false negatives or false positives of up to 0.1. However, the SBCE probability estimation is not robust to colonization probability heterogeneity between patches. When we applied the SBCE probability method to the real monitoring dataset, we found a weak influence of the flowering months of the species and of the green space the closest to the metapopulation on SBCE probability estimates.

4. The study suggests that the measurement of seed bank contribution is less data-demanding than assessment of seed bank presence. Applying the es-

51 timination method to the monitoring of tree bases plant species highlights a
52 significant contribution of the seed bank to plant metapopulation dynamics
53 in an urban environment, and illustrates how the method can be applied on
54 real-world datasets.

55 **Keywords :** Hidden Markov Model, metapopulation, plants, Propagule Rain model,
56 seed bank, urban biodiversity, tree bases

Introduction

An important issue in ecology and conservation biology is determining the mechanisms underlying persistence of plant or animal populations in fragmented landscapes (Fahrig, 2003). In plants, the seed bank, i.e the spontaneous storage of seeds within the soil, plays a critical role in metapopulation and community dynamics (Fenner, 2017). However, assessing directly the presence of a seed bank, for instance by putting soil samples in germination chambers, and measuring its associated parameters is challenging. Therefore conceptual approaches and statistical tools allowing one to estimate these quantities using widespread data such as presence/absence data can prove very useful. One suitable conceptual framework for studying patchy environments is the metapopulation theory, first introduced in (Levins, 1969). A metapopulation is defined as a population living in a set of patches that can be colonized or go extinct, the regional persistence of the species resulting from a balance between local colonizations and extinctions (MacArthur & Wilson, 1967). Statistical tools have been developed to allow parameter inference for a broad range of metapopulation models (see e.g (Moilanen, 1999, 2004)), and were fruitfully used in studies on insects (Hanski, 2011; Moilanen, Smith, & Hanski, 1998) or small mammals (Ozgul, Armitage, Blumstein, & Oli, 2006).

As plants form populations with a strong spatial structure and can only move from one patch to another as propagules, metapopulation models appear at first as particularly suited to their study (Husband & Barrett, 1996). Yet classical metapopulation models do not account for seed banks, which are common in seed plants (Baskin & Baskin, 2014), potentially leading to erroneous estimates of extinction and colonization rates (Fréville, Choquet, Pradel, & Cheptou, 2013) and making these models generally irrelevant for studying plant metapopulation dynamics (Freckleton & Watkinson, 2002). New models taking into account the influence of a seed bank were developed recently (Fréville et al., 2013; Borgy, Reboud, Peyrard, Sabbadin, & Gaba, 2015). These models consider the seed bank state as an *hidden state*, which is not visible but which influences patch occupancy, and which can be estimated from patch occupancy data. In this article, we elaborated a new method to characterize seed bank contribution to metapopulation dynamics based

on the model introduced in (Pluntz et al., 2018). This model allows parameter inference on a variant with a seed bank of a classical model, the *Propagule Rain Model* (or PRM) (Gotelli, 1991), in which patches are colonized or go extinct independently from each other, with the same colonization or extinction probability. However, a limit of the model proposed by (Pluntz et al., 2018) is that it would not be applicable in many real-world situations. It indeed requires either a long monitoring duration or several thousand patches to be monitored in order to accurately estimate all parameters (that is, germination, colonization and seed bank death probabilities). To overcome this problem, we introduced in the present study a new metric more accurately estimated than the metapopulation parameters for identical monitoring duration and number of patches, providing information about the influence of the seed bank on plant population dynamics in real populations. Then, we used this metric with both theoretical and real metapopulation data. We called this metric the *seed bank characteristic event* probability (SBCE probability). As it only measures the contribution of the seed bank to the observed standing vegetation dynamics, it is less informative than knowing all seed bank parameters, but the goal of this study was to show that it is accurate in more real-life situations. We performed analyses based on simulated presence/absence time series data, and on time series data in which we introduced some flaws commonly found in real datasets, in order to give bounds on the number of patches or of years of observation needed to fulfill different accuracy requirements. Our goal was to provide guidelines on how to design the data collection step.

As a case study, we used the estimation method on annual floristic inventories of natural and spontaneous flora carried out from 2009 to 2018 on 1324 tree bases located in Paris, France (the *Paris 12* dataset). Indeed, the population of plants in urban tree bases is located inside an inhospitable matrix and has an high turnover, which makes metapopulation models particularly suited (Dornier, Pons, & Cheptou, 2011). Studies using presence/absence data for various species present in urban tree bases considered as metapopulations were already carried out, but to our knowledge none accounted for seed bank potential presence in tree bases (Omar et al., 2019; Dornier et al., 2011). We

also attempted to relate the contribution of the seed bank to the observed population dynamics to species traits and environmental characteristics. The results are intended to increase the understanding of population dynamics inside an urban environment.

Overall, our study (i) gives insights on the importance of seed bank contribution to plant metapopulation dynamics within an urban environment, and (ii) provides a comprehensive framework to detect the effects of seed banks in plant metapopulations, which can be applied on a wide range of ecological systems, including but not restricted to urban environments.

Material and methods

Model used

The model from (Pluntz et al., 2018) we used in this study is a variant of the Propagule Rain Model (PRM) (Gotelli, 1991). In the PRM, colonization and extinction probabilities do not depend on the current state of the metapopulation, and are constant over patches and time. A seed bank can be introduced using Hidden Markov Model (HMM) techniques (Cappé, Moulines, & Rydén, 2005; Rabiner, 1989). The seed bank contains the seeds that were just produced by standing vegetation of the focal patch or came from colonization events by the propagule rain, along with seeds produced by previous generations that did not germinate yet and are still alive. Since all plants originate from the seed bank of the patch they are in, the presence of plants at one time step means that the seed bank contained seeds just before germination could occur. Therefore metapopulation parameters can be estimated along with the presence/absence of seeds in the seed bank at each time-step.

The model is characterized by these three parameters :

- the joint probability of seed germination and of survival of seedlings until adulthood.

This parameter will be called *germination probability*, and denoted g , following the existing literature.

- the *colonization probability* c of the patch by external seeds entering the seed bank.
- the *seed bank extinction probability conditional on the seed bank not having germinated* d . This parameter is the probability that the seed bank will not survive until the next generation, assuming it has not germinated yet.

The initial proportion p_0 of patches containing a seed bank can be considered as an extra parameter of the model.

The model evolves as follows : for each patch, if the seed bank is not empty, seeds can germinate with probability g . If they do, the plants will grow and produce seeds which will refill the seed bank. Otherwise, the seed bank can survive until the next generation with probability $1-d$. New seeds can enter the seed bank during a colonization event with probability c independently of the presence of standing vegetation or seeds in any given patch (see figure 1). The main difference with the PRM model is that if the seed bank death probability d is strictly less than 1, then germination can be delayed. Hereafter, models for which $d < 1$ will be denoted as SB+ (*Seed-Bank Plus*) and those for which $d = 1$, corresponding to the classical PRM, will be denoted SB- (*Seed-Bank Minus*).

The procedure for estimating parameters uses the Expectation-Maximization (EM) algorithm (Dempster, Laird, & Rubin, 1977) in order to find the best parameter fit (see details in (Pluntz et al., 2018)). We set the number of iterations in the algorithm to 200 in order to ensure convergence of parameters. As the EM algorithm can converge to a local but non-global maximum, the choice of the initial conditions can affect the value returned by the algorithm. We preliminary checked with simulated datasets that convergence of the EM algorithm to a non-global maximum was rather unlikely (results not shown).

The algorithm was first used to get theoretical results on the performances of the SBCE estimation method. We simulated datasets for different parameter values for g , c , d and p_0 and applied the SBCE estimation method on these simulated datasets in order to identify conditions under which the SBCE probability offers reliable estimates using a set of performance criteria. Then, we used the estimation method on a real monitoring dataset, the Paris 12 dataset (see below). Due to the structure of this dataset, we implemented a

variant of the estimation method supporting one year of missing data.

Criterion for seed bank identification : SBCE probability

Seed Bank Characteristic Event (SBCE) probability, denoted $\mathbb{P}_{SBCE}$, is defined as the probability for standing flora to produce seeds that (1) will stay in the seed bank during more than one year (without germinating nor dying), and (2) will germinate before new seeds come from an external source. In other words, $\mathbb{P}_{SBCE}$ is the probability that standing flora in a given patch at a given time-step contributes with delay to population dynamics, and that this contribution is not hidden by colonization events happening in between. This probability can be computed knowing g , c and d . Denoting SB_t the event *during t years, the seed bank does not germinate, nor receive new seeds from external source, nor dies*, then :

$$\begin{aligned}\mathbb{P}_{SBCE} &= \sum_{t \geq 1} \mathbb{P}(SB_t) \times g \\ &= g \times \sum_{t \geq 1} \mathbb{P}(SB_1)^t \\ &= g \times \sum_{t \geq 1} [(1-g)(1-c)(1-d)]^t \\ &= g \times \frac{(1-g)(1-c)(1-d)}{1 - (1-g)(1-c)(1-d)}\end{aligned}$$

SBCE probability can be interpreted as a measure of seed bank contribution to metapopulation dynamics : it takes into account that some parameter values, e.g high colonization or seed bank extinction probabilities, make it hard for delayed germination events to be observed. As a result it cannot be used per se to conclude to seed bank presence or absence, but is a proxy of seed bank contribution to the metapopulation dynamics. A consequence is that low SBCE probability does not mean that there is no seed bank, since a high colonization probability yields comparable effects.

Hereafter, we will consider the effect of the seed bank to be :

- *weak* if SBCE probability is lower than 0.05

190 • *medium* if SBCE probability is in the interval $[0.05, 0.10]$

191 • *strong* if SBCE probability is higher than 0.10.

192 These thresholds were defined arbitrarily using values taken by the SBCE probability in
193 situations where the observed effects of the seed bank can be considered weak, strong or
194 intermediate. Continuous values taken by the SBCE probability for parameter values in
195 table 1 are presented in Supplementary Information (A.1). An estimator of SBCE proba-
196 bility must satisfy two requirements so that it can accurately show seed bank contribution
197 to metapopulation dynamics. First, the estimated SBCE probability must be accurately
198 estimated for datasets generated with a SB+ model. We considered the estimation was
199 correct if the root-mean-square error (RMSE) on SBCE probability estimation was lower
200 than the threshold of 0.14 used in (Pluntz et al., 2018). Then, it must avoid, as much as
201 possible, identification of medium or strong seed bank effects in datasets generated with
202 a SB- model, for which $\mathbb{P}_{SBCE} = 0$. In this article, the rate of false identification of a
203 medium to strong seed bank effect in a model without seed bank (SB-), denoted $p_{falseSB}$
204 thereafter, is deemed satisfying if it is below 0.05.

205 In order to assess the performance of the criterium, we investigated the sensibility of
206 $\mathbb{P}_{SBCE}$ estimation to c , g , d and p_0 by building 24 distinct parameter sets combining a
207 broad range of values for these parameters (16 corresponding to SB+ models, and 8 to
208 SB- models, see table 1). Each parameter set was used to generate 30 datasets for each
209 of the 4 monitoring durations and 8 numbers of patches listed in table 1. We chose to
210 consider durations of at most 20 years and at most 1000 patches.

211 We computed the estimated SBCE probability by performing parameter fits of c , g , d
212 and p_0 on the simulated datasets. The accuracy of the SBCE probability estimation was
213 tested separately for each SB+ parameter set, time duration and number of patches by
214 computing the RMSE. We determined the minimal number of patches needed to sat-
215 isfy a RMSE threshold of 0.14 (as in (Pluntz et al., 2018)). For SB- parameter sets, as
216 $\mathbb{P}_{SBCE} = 0$, we instead computed $p_{falseSB}$ for each combination of monitoring duration
217 and number of patches.

218 The performance of the criterium were compared to the one of a criterium comparing

the AIC of parameter fits with or without a constraint on d (i.e, comparing the AIC of parameter fits on a SB+ and a SB- model). The AIC-based criterium was deemed satisfying if it was able to identify the correct model at least 25 times out of 30 (i.e, 83% of correct assignments).

Testing SBCE estimation robustness

Before applying the parameter estimation method to real data, we tested the algorithm robustness to several flaws commonly found in real datasets : false negatives, false positives and parameter heterogeneity.

Parameter heterogeneity was only studied for colonization probability by assuming that a proportion of patches has a colonization probability equal to 0. Details of the protocols are provided in Supporting Information (A.2).

Applying SBCE probability to real-world monitoring data

The real dataset used in this study to apply the SBCE probability, Paris 12, consists of floristic inventories of 1324 tree bases located in Paris 12th administrative district, carried out annually between 2009 and 2018. Natural and spontaneous flora was inventoried exhaustively over the entire period, except in 2013, when a limited number of species were tracked. For those species that were not monitored in 2013, we used the missing data variant of the model for the year 2013. The taxonomic reference is the French Flora Reference TAXREF v8.0 (Gargominy et al., 2014). See Supporting Information (B.1) or (Omar, Al Sayed, Barré, Halwani, & Machon, 2018) for information about nomenclature and the species and streets monitored.

For almost every species, a high proportion of patches were never occupied from 2009 to 2018. We interpreted this as being due to colonization heterogeneity, and considered that patches which were never colonized had a colonization probability equal to 0. Since the germination and seed bank death probabilities cannot be estimated for a group of patches that were never occupied, we removed these patches from the analysis.

We considered that the tree bases present in different streets represented distinct metapop-

246 ulations. Moreover, we considered that the metapopulation parameters could be different
 247 from one species to another, and for a given species, from one street (i.e one metapop-
 248 ulation) to another. Each pair of species and street was analyzed separately, and was
 249 retained only if the species was observed at least once in more than 20 distinct tree bases
 250 of the street (the size of a street ranged from 31 to 186 tree bases). For each pair, seed
 251 bank contribution was assessed under the hypothesis that the species' patch occupancy
 252 dynamics followed the Propagule Rain Model. Parameter estimation for SB+ model was
 253 carried out using the missing data variant when the species was not inventoried in 2013.
 254 We then tested whether the estimated SBCE probability was affected by the nature of the
 255 closest green space (see (Omar et al., 2019)), the seed dispersal mechanism, the flowering
 256 months (extracted from the database of the collaborative network of French botanists
 257 "Tela botanica" (<http://www.tela-botanica.org>), the releasing height of the seeds and the
 258 seed weight (mean value obtained from the LEDA database (Kleyer et al., 2008)). Be-
 259 cause almost all species flower in summer months (June, July and August) and due to
 260 high correlation in flowering for consecutive months, we limited the flowering period data
 261 to two binary variables : early flowering (months of March, April or May) and late flow-
 262 ering (September or October). We used mixed-effect linear regression model with SBCE
 263 probability as the response variable, species and street identity as random factors, the
 264 greenspace the closest to the street among 4 (one park, one footpath, railways and the
 265 Seine river) and the dispersal mechanism (anemochorous, autochorous, barochorous or
 266 epizoochorous) as fixed qualitative variables, the releasing height of the seeds and the seed
 267 weight as fixed quantitative variables and the early and late flowering as binary variables,
 268 assuming a Gaussian distribution. We checked whether the requirements of independence
 269 of residuals, normality of residuals and homogeneity of variances were met. The model
 270 was implemented with the lmer function of the lme4 R package (Bates, Mächler, Bolker,
 271 & Walker, 2014). Besides the above regression model, we also performed repeatability
 272 analyses to provide an overview of the variation of the germination probability and SBCE
 273 probability among streets and among species. The repeatability analysis was based on
 274 1000 parametric bootstraps as implemented in the rptR package of R (Stoffel, Nakagawa,

& Schielzeth, 2017). The statistical significance of the repeatability of each metric was tested by a likelihood ratio test comparing the model fit of a model including a grouping factor (here, the species or the street) and one excluding it.

Results

Criterion for seed bank identification

For each monitoring duration, we could find a number of patches ensuring the criterion fulfillment ($RMSE < 0.14$ on SBCE probability estimation for SB+ models, and $p_{falseSB} < 0.05$ for SB- models). No such number of patches could be found for the criterion comparing AIC of parameter fits on a model with or without seed bank, which is not based on SBCE probability (see Supporting Information (A.3)).

Overall, the minimal number of patches required decreased from 500 for a monitoring lasting 3 years to 30 for a monitoring lasting 20 years, and was higher for SB+ models than for SB- models. Complete results can be found in Supporting Information (A.1).

Testing SBCE estimation robustness

Overall, the introduction of false negatives, false positives or heterogeneous colonization increased RMSE on SBCE probability estimation for SB+ models, and $p_{falseSB}$ for SB- models.

For SB+ models, for all the false positive or false negative rates considered, the RMSE almost always stayed below the 0.14 threshold. The introduction of heterogeneous colonization by making some patches not colonizable led to the threshold being quickly exceeded, except for 500 patches and 3 years of monitoring.

For SB- models, when we introduced false negatives or false positives, increasing the monitoring duration or the number of patches led to a decrease of $p_{falseSB}$ (albeit less marked for a false negative rate of 0.2), and the 0.05 threshold could be met for a sufficient monitoring duration or number of patches, except for a false negative rate of 0.2 (see table 3). Conversely, when we introduced heterogeneous colonization (i.e, when we set

the colonization probability of some patches to 0), increasing the number of patches had no marked effect, and increasing the monitoring duration actually made $p_{falseSB}$ increase, reaching values of up to 0.5. This effect was particularly marked when considering only the false identification rate of a strong (instead of a medium to strong) seed bank effect. Consequently, no combination of number of years of observation and number of patches monitored fulfilled the accuracy requirements, even for a proportion of non-colonizable patches of 0.05.

Complete results can be found in Supporting Information (A.4).

Applying SBCE probability to real-world monitoring data

The analysis highlighted a high variability of SBCE probabilities, both between species and within species. A medium to strong seed bank effect was detected in at least one street for most species. We further investigated the reason for this variation looking at estimates of the germination probability g . Germination probabilities were generally fairly low, most of the time below 0.5, no matter the species or the street. The repeatability analysis indicated that the germination and SBCE probabilities were very consistent between streets for a given species (repeatability $R = 0.595 \pm 0.07$ for germination and $R = 0.38 \pm 0.08$ for SBCE, p-value $< 10^{-4}$ for both repeatabilities). In contrast, repeatabilities of these probabilities between species for a given street were low ($R = 0.058 \pm 0.03$, p-value = 7×10^{-4} for germination and $R = 0.058 \pm 0.04$, p-value = 0.003 for SBCE). Overall, these results suggest that both germination and SBCE probabilities primarily depend on species rather than location. Results are summarized in figures 2 and 3. Estimations of SBCE and germination probabilities for each pair of species and street (i.e, each metapopulation) analysed can be found in Supporting Information (B.2). According to the regression model, none of our explaining variables was significantly correlated with the SBCE probability (see detailed regression results in Supporting Information (B.2)). However, the type of the nearest green space and the flowering period exhibited marginally significant effects : late flowering and being closest to the park were associated with higher SBCE probability (figure 3).

Discussion

In this paper, we propose a new metric, the seed bank characteristic event probability (SBCE) providing information on the contribution of the seed bank to the observed standing vegetation dynamics in plant metapopulations. Our results indicated that the SBCE performs well in a wide range of situations and provided evidence of a significant contribution of the seed bank to plant metapopulation dynamics in an urban environment. In biology, as in other disciplines, it is sometimes more practical and straightforward to make inferences about a process within a system by observing the effects of that process on the system rather than the process itself. This approach is central to the study of metapopulations, where a process such as dispersal is often studied indirectly by examining its consequences in terms of genetic structuring or recolonization dynamics. This idea is also at the core of the analytical framework recently developed (Fréville et al., 2013; Borgy et al., 2015) to study seed banks, which are very difficult to detect on a large scale, but whose consequences in terms of plant metapopulation dynamics can be crucial (Fenner, 2017).

Theoretical analysis

Our analysis indicates that the SBCE approach is relevant to evaluate the contribution of a potential seed bank to the dynamics of a metapopulation in which patches are independent, i.e., in which the colonization or extinction of a patch does not depend on the presence of the species in the other patches, even when the number of monitored patches and monitoring duration are limited. In particular, we showed that for limited monitoring durations and number of patches, the SBCE probability provides more precise predictions than model identification methods using AIC and based on estimates of the germination probability g , the colonization probability c and the seed bank death probability d . This could be expected, since even though the SBCE probability is computed using g , c and d , it depends mostly on g , which according to (Pluntz et al., 2018) is well estimated. Conversely, model estimation methods depend more on estimates of d , which

is more difficult to estimate. For a sufficient but still achievable number of patches or of years of observations, the SBCE probability was well estimated when seed bank existed and correctly reflected the absence of a seed bank otherwise (see Supporting Information (A.1)). As this metric measures the contribution of the seed bank to the global metapopulation dynamics, it gives information on whether the seed bank is essential to the observed dynamics. However, it is not suited for getting estimates of the seed bank parameters, namely probabilities of germination and survival of seeds (but see (Pluntz et al., 2018)).

Our work highlights that false negatives or false positives in standing vegetation detection and heterogeneous colonization have contrasted impacts on the estimation of the SBCE probability. On the one hand, for models without seed bank, increasing false positive rate or false negative rate lead to an increase of the rate of false identification of a medium to strong seed bank effect $p_{falseSB}$. The presence of false positives can be mitigated by monitoring more patches over a longer duration, while the presence of false negatives can be mitigated only when the false negative rate is not too high. This implies that monitoring methods reducing the false negative rate may increase the accuracy of the SBCE probability estimate, even if they are traded-off by an increase of the rate of false positives. On the other hand, with strongly heterogeneous colonization (e.g., in the special case in which a fraction of patches can never be colonized), the SBCE probability cannot be well estimated, no matter the monitoring duration or the number of patches. Therefore it is crucial to identify heterogeneous colonization situations, whose impact on estimation accuracy cannot be mitigated by improving the monitoring effort. Other statistical methods can be used in order to identify heterogeneous colonization cases, allowing to treat the case where colonization probabilities are heterogeneous but potentially all non-zero, for instance using mixture models (Robin, 2018). However these methods are far more computationally intensive than the simple one we used. The consistency of estimates of the probability of germination g between streets for most species in the *Paris 12* dataset suggests that the method we introduced is still efficient.

The fact that increasing the number of patches can mitigate the impact of false negatives

or false positives suggests that the method may be used on data coming from citizen science programs, for which the number of patches monitored can be very large. Recent empirical results on citizen science programs showed that, with a standardized protocol and good training methods, error rates in datasets can reach values lower than 0.05 (Fuccillo, Crimmins, de Rivera, & Elder, 2015; Ratnieks et al., 2016), which makes these programs compatible with our SBCE approach.

The Paris 12 dataset

The analysis performed on the *Paris 12* dataset of plants present in tree bases in Paris showed that 55 % of the pairs of species and streets analysed exhibited a medium to strong seed bank effect, suggesting that seed banks have a key influence on plant metapopulation dynamics in this type of urban environment. Moreover, estimates of both germination probability g and SBCE probability are consistent between streets for a given species, but they are not consistent from one species to another for a given street. Besides, the spatial variation of SBCE estimates among streets is weakly related to which green space is the closest to the street. Explaining this pattern is difficult, as it may result from the green space characteristics, from the design of the tree bases or the tree-base management performed by the green space services of the city of Paris. Most estimates of g are low, below 0.5 or even 0.3, which means that either germination per se is low, either survival of seedlings until adulthood is low. Potential explanations are a strong competition between species, or a removal of plants before seed production by gardeners.

Our analysis does not uncover any influence of the dispersal mechanism on the SBCE probability. This result is in line with the findings of (Omar et al., 2018), who showed that the distribution of species in Paris tree bases was not correlated to the dispersal mechanism nor to the weight of the seeds, and hypothesised that this was partly due to human activity spreading all seeds no matter the weight or the dispersal device (Sukopp, 2004; Von der Lippe & Kowarik, 2007).

Our study also highlights a weak influence of the flowering period on the SBCE probability. Species whose flowering period starts early (before May) exhibited a higher mean

413 estimated SBCE probability than those whose flowering period starts later. One possi-
 414 ble interpretation is that plant species that flower early are more detectable than those
 415 flowering late, so they are more likely to be removed by gardeners, which affects the
 416 probability of survival of seedlings until adulthood and the colonization probability.

417 The applicability of the SBCE probability on the *Paris 12* dataset has some limitations.
 418 First, according to our theoretical analysis, estimations of the SBCE probability for a
 419 monitoring duration of 10 years is expected to be accurate if at least 100 tree base sare
 420 considered, and more in case of false negatives or false positives. As streets in our dataset
 421 have between 30 and 150 tree bases and as some tree bases were removed from analysis
 422 to avoid issues related to heterogeneous colonization, small errors on SBCE probability
 423 estimates have to be expected, and particularly a slight overestimation of SBCE proba-
 424 bilities when no seed bank is present (i.e when SBCE probability is equal to 0).

425 Second, we assumed that tree bases' dynamics are independent of each other (i.e., no
 426 colonization from one patch to another). As a result, we did not consider the alternative
 427 hypothesis of Levins metapopulation model ([Levins, 1969](#)) with a seed bank. In this
 428 model, colonization events do not bring seeds from a propagule rain, but from neigh-
 429 bouring patches. Therefore, colonization probability is not the same for all patches,
 430 and depends on the state of the metapopulation. An estimation method for this model
 431 exists ([Le Coz, Cheptou, & Peyrard, 2019](#)), but uses abundance data instead of pres-
 432 ence/absence data, and would need to be adapted to handle missing data. Moreover, the
 433 fact that we recovered results previously documented with a different approach and the
 434 consistency of g estimates for a given species between streets suggests that the assump-
 435 tion of independent patches was correct.

436 Our results show that measuring seed bank contribution to plant metapopulation dynam-
 437 ics is less data-demanding than assessing seed bank presence, while being robust to the
 438 presence of false negatives or false positives. Our method can be applied to a wide range
 439 of urban and non-urban metapopulations, and can be used on datasets collected using cit-
 440 izen science in order to increase substantially the understanding of plant metapopulation
 441 dynamics.

Acknowledgments

We wish to thank Mona Omar, Marion Noualhaguet, Marion Dubois, Ambre Zéléla Bouvard, Paul Haenel, Noëlie Maurel, Florence Devers, Hélène Beaugeard, Sébastien Julliard, Laure Schneider-Maunoury and Gwendoline Chastel who carried out the floristic survey from 2009 to 2018. This work was supported by ANR (Agence Nationale de la Recherche) Écoville [ANR 14 CE22-0021] and was done during the internship of AL at the CESCO laboratory as a student of ENS Paris.

Authors' contributions

All authors conceived the ideas and AL designed methodology. AL analyzed the data and led the writing of the manuscript. All authors contributed critically to the drafts and gave final approval for publication.

Data availability

The *Paris 12* dataset is available on Zenodo (<http://doi.org/10.5281/zenodo.3770339>).

References

- Baskin, C., & Baskin, J. (2014). *Seeds : Ecology, biogeography, and evolution of dormancy and germination*. Academic Press, San Diego.
- Bates, D., Mächler, M., Bolker, B., & Walker, S. (2014). Fitting linear mixed-effects models using lme4. *arXiv preprint arXiv:1406.5823*.
- Borgy, B., Reboud, X., Peyrard, N., Sabbadin, R., & Gaba, S. (2015). Dynamics of weeds in the soil seed bank: a hidden markov model to estimate life history traits from standing plant time series. *PloS one*, 10(10), e0139278. doi: [10.1371/journal.pone.0139278](https://doi.org/10.1371/journal.pone.0139278)
- Cappé, O., Moulines, E., & Rydén, T. (2005). *Inference in hidden markov models*. Springer-Verlag, New York.
- Dempster, A. P., Laird, N. M., & Rubin, D. B. (1977). Maximum likelihood from incomplete data via the em algorithm. *Journal of the royal statistical society. Series B (methodological)*, 1–38. doi: [10.1111/j.2517-6161.1977.tb01600.x](https://doi.org/10.1111/j.2517-6161.1977.tb01600.x)
- Dornier, A., Pons, V., & Cheptou, P.-O. (2011). Colonization and extinction dynamics of an annual plant metapopulation in an urban environment. *Oikos*, 120(8), 1240–1246. doi: [10.1111/j.1600-0706.2010.18959.x](https://doi.org/10.1111/j.1600-0706.2010.18959.x)
- Fahrig, L. (2003). Effects of habitat fragmentation on biodiversity. *Annual review of ecology, evolution, and systematics*, 34(1), 487–515. doi: [10.1146/annurev.ecolsys.34.011802.132419](https://doi.org/10.1146/annurev.ecolsys.34.011802.132419)
- Fenner, M. (2017). Ecology of seed banks. In *Seed development and germination* (pp. 507–528). Routledge.
- Freckleton, R., & Watkinson, A. (2002). Large-scale spatial dynamics of plants: metapopulations, regional ensembles and patchy populations. *Journal of Ecology*, 90(3), 419–434. doi: [10.1046/j.1365-2745.2002.00692.x](https://doi.org/10.1046/j.1365-2745.2002.00692.x)
- Fréville, H., Choquet, R., Pradel, R., & Cheptou, P.-O. (2013). Inferring seed bank from hidden markov models: new insights into metapopulation dynamics in plants. *Journal of Ecology*, 101(6), 1572–1580. doi: [10.1111/1365-2745.12141](https://doi.org/10.1111/1365-2745.12141)
- Fuccillo, K. K., Crimmins, T. M., de Rivera, C. E., & Elder, T. S. (2015). Assessing

accuracy in citizen science-based plant phenology monitoring. *International journal of biometeorology*, 59(7), 917–926. doi: [10.1007/s00484-014-0892-7](https://doi.org/10.1007/s00484-014-0892-7)

Gargominy, O., Tercerie, S., Régnier, C., Ramage, T., Schoelinck, C., Dupont, P., ... Poncet, L. (2014). Taxref v8. 0, référentiel taxonomique pour la france: Méthodologie, mise en oeuvre et diffusion. *Rapport SPN*, 42, 2014.

Gotelli, N. J. (1991). Metapopulation models: the rescue effect, the propagule rain, and the core-satellite hypothesis. *The American Naturalist*, 138(3), 768–776. doi: [10.1086/285249](https://doi.org/10.1086/285249)

Hanski, I. A. (2011). Eco-evolutionary spatial dynamics in the glanville fritillary butterfly. *Proceedings of the National Academy of Sciences*, 108(35), 14397–14404. doi: <https://doi.org/10.1073/pnas.1110020108>

Husband, B. C., & Barrett, S. C. (1996). A metapopulation perspective in plant population biology. *Journal of Ecology*, 461–469. doi: [10.2307/2261207](https://doi.org/10.2307/2261207)

Kleyer, M., Bekker, R., Knevel, I., Bakker, J., Thompson, K., Sonnenschein, M., ... others (2008). The leda traitbase: a database of life-history traits of the north-west european flora. *Journal of Ecology*, 96(6), 1266–1274. doi: [10.1111/j.1365-2745.2008.01430.x](https://doi.org/10.1111/j.1365-2745.2008.01430.x)

Le Coz, S., Cheptou, P.-O., & Peyrard, N. (2019). A spatial markovian framework for estimating regional and local dynamics of annual plants with dormancy. *Theoretical population biology*, 127, 120–132. doi: [10.1016/j.tpb.2019.03.002](https://doi.org/10.1016/j.tpb.2019.03.002)

Levins, R. (1969). Some demographic and genetic consequences of environmental heterogeneity for biological control. *American Entomologist*, 15(3), 237–240. doi: [10.1093/besa/15.3.237](https://doi.org/10.1093/besa/15.3.237)

MacArthur, R. H., & Wilson, E. O. (1967). *The theory of island biogeography*. Princeton university press, New Jersey.

Moilanen, A. (1999). Patch occupancy models of metapopulation dynamics: efficient parameter estimation using implicit statistical inference. *Ecology*, 80(3), 1031–1043. doi: [10.1890/0012-9658\(1999\)080\[1031:POMOMD\]2.0.CO;2](https://doi.org/10.1890/0012-9658(1999)080[1031:POMOMD]2.0.CO;2)

Moilanen, A. (2004). Spomsim: software for stochastic patch occupancy mod-

els of metapopulation dynamics. *Ecological modelling*, 179(4), 533–550. doi:
10.1016/j.ecolmodel.2004.04.019

Moilanen, A., Smith, A. T., & Hanski, I. (1998). Long-term dynamics in a metapopulation of the american pika. *The American Naturalist*, 152(4), 530–542. doi:
10.1086/286188

Omar, M., Al Sayed, N., Barré, K., Halwani, J., & Machon, N. (2018). Drivers of the distribution of spontaneous plant communities and species within urban tree bases. *Urban Forestry & Urban Greening*, 35, 174–191. doi: 10.1016/j.ufug.2018.08.018

Omar, M., Schneider-Maunoury, L., Barré, K., Al Sayed, N., Halwani, J., & Machon, N. (2019). Colonization and extinction dynamics among the plant species at tree bases in paris (france). *Ecology and Evolution*, 9(15), 8414–8428. doi: 10.1002/ece3.4954

Ozgul, A., Armitage, K. B., Blumstein, D. T., & Oli, M. K. (2006). Spatiotemporal variation in survival rates: implications for population dynamics of yellow-bellied marmots. *Ecology*, 87(4), 1027–1037. doi: 10.1890/0012-9658(2006)87[1027:SVISRI]2.0.CO;2

Pluntz, M., Coz, S. L., Peyrard, N., Pradel, R., Choquet, R., & Cheptou, P.-O. (2018). A general method for estimating seed dormancy and colonisation in annual plants from the observation of existing flora. *Ecology letters*. doi: 10.1111/ele.13097

Rabiner, L. R. (1989). A tutorial on hidden markov models and selected applications in speech recognition. *Proceedings of the IEEE*, 77(2), 257–286. doi: 10.1109/5.18626

Ratnieks, F. L., Schrell, F., Sheppard, R. C., Brown, E., Bristow, O. E., & Garbuzov, M. (2016). Data reliability in citizen science: learning curve and the effects of training method, volunteer background and experience on identification accuracy of insects visiting ivy flowers. *Methods in Ecology and Evolution*, 7(10), 1226–1235. doi: 10.1111/2041-210X.12581

Robin, S. (2018, January). *Models with hidden structure with applications in biology and genomics*. Master 2 Mathématiques pour les Sciences du Vivant, Université Paris-Sud.

Stoffel, M. A., Nakagawa, S., & Schielzeth, H. (2017). rptR: Repeatability estimation

542 and variance decomposition by generalized linear mixed-effects models. *Methods in*
543 *Ecology and Evolution*, 8(11), 1639–1644.

544 Sukopp, H. (2004). Human-caused impact on preserved vegetation. *Landscape and urban*
545 *planning*, 68(4), 347–355. doi: [10.1016/S0169-2046\(03\)00152-X](https://doi.org/10.1016/S0169-2046(03)00152-X)

546 Von der Lippe, M., & Kowarik, I. (2007). Long-distance dispersal of plants by vehi-
547 cles as a driver of plant invasions. *Conservation Biology*, 21(4), 986–996. doi:
548 [10.1111/j.1523-1739.2007.00722.x](https://doi.org/10.1111/j.1523-1739.2007.00722.x)

Figure 1: Graphical representation of the variant with seed bank of the Propagule Rain Model. White arrows indicate transitions that always occur, while black arrows indicate transitions that only occur with a fixed probability.

Parameter and notation	SB+ model	SB- model
Colonization probability (c)	0.3, 0.7	0.3, 0.7
Germination probability (g)	0.3, 0.7	0.3, 0.7
Seed bank death probability (d)	0.2, 0.6	1
Initial proportion of seeds (p_0)	0.3, 0.7	0.3, 0.7
Years of observation	3, 5, 10, 20	3, 5, 10, 20
Number of patches	10, 30, 50, 100, 200, 500, 800, 1000	10, 30, 50, 100, 200, 500, 800, 1000

Table 1: Parameter sets and test conditions used for testing the seed bank identification criterium.

(a) SBCE probability estimates

(b) Germination probability estimates

Figure 2: (a) Distribution of SBCE probability estimates for species present in at least 20 tree bases of at least 3 different streets. (b) Distribution of germination probability estimates for species present in at least 20 tree bases of at least 3 different streets. Species flowering early are indicated in turquoise, and species flowering late in white. Species whose flowering period could not be found on Tela Botanica are not indicated.

Monitoring duration	3 years		5 years		10 years		20 years	
Model	SB+	SB-	SB+	SB-	SB+	SB-	SB+	SB-
Minimal number of patches	500	200	200	100	100	100	30	30

Table 2: Minimal number of patches needed to achieve $RMSE < 0.14$ on SBCE probability estimation for SB+ models, and $p_{falseSB} < 0.05$ for SB- models, depending on monitoring duration.

Monitoring duration	3 years	5 years	10 years	20 years
False negative rate	Number of patches			
0	500	200	100	30
0.05	1000	500	500	200
0.1	X	1000	500	500
0.2	X	X	X	X
False positive rate	Number of patches			
0	500	200	100	30
0.05	X	1000	500	100
0.1	X	X	500	500
0.2	X	X	X	800

Table 3: Minimal number of patches needed to satisfy both $RMSE < 0.14$ on SBCE probability estimation for SB+ models and $p_{falseSB} < 0.05$, after having introduced false negatives or false positives in the dataset, for different monitoring durations and rates of false negatives or false positives. The symbol X means that criterium satisfaction could not be met.

Figure 3: (a) Distribution of SBCE probability estimates in streets classified by the nature of the closest green space (see Supplementary information (B.1) for green spaces names). (b) Distribution of SBCE probability estimates for species with early flowering period (March, April or May), compared to those with late flowering period.