Ecography

ECOG-02575
Barbu, C. M., Sethuraman, K., Billig, E. M. W. and Levy, M. Z. 2017. Two-scale dispersal estimation for biological invasions via synthetic likelihood. – Ecography doi: 10.1111/ecog.02575

Supplementary material

Supplementary materials: Appendix 1-6 to Two-Scale Dispersal Estimation for Biological Invasions via Synthetic Likelihood

Corentin M. Barbu *et al.*

Annexe 1 Complexity when defining an analytical likelihood for multi-scaled processes

The hop-jump model described in the main text exemplifies how difficulties in defining an analytical likelihood can arise from a simple model describing several scales of dispersal. Because observation time points are fairly distant, a number of successive steps (here, hops or jumps) from initial infested locations to final infested locations could have occurred between the observations.

For example in Figure 1, if we considered that only one step occurred between the two observations, the whole group of newly infested households at the bottom right of the village in the final observations would have to result from many independent long range dispersals. Assuming only one time step between the two observations has two important effects on parameter estimation:

1. the rate of long range dispersal tends to be overestimated;
2. the credible intervals tend to be too narrow as each infestation is incorrectly considered as an independent unit of information on dispersal.

In the synthetic likelihood approach, by simulating and numerically estimating a likelihood of the general resulting pattern, we bypass these difficulties arising from unobserved chains of colonization between the two observed time points.

If information on time since infestation was available, methods integrating over possible paths of infestation using analytical likelihoods and reversible jump MCMC might be feasible. However, given the binary nature of the data we characterize, such methods are computationally challenging. Hereafter we show how an analytical likelihood quickly becomes untractable as we consider an increasing number of possible steps between the first and the second observations.

Annexe 1.1 Considering only one step

Be X_{ti} the binary value for infestation of unit i at time t. We have two observation times 0 and 1. The likelihood of the model can easily be defined for our model if we consider that only one step can occur between the two observations. The state in each unit at a given time is independent of the state in the other units at the same time as it only depends on the state at the previous time. This is a first order approximation of the analytical likelihood.

$$L(X) = P(X_1|X_0) = \prod_i P(X_{1i}|X_0)$$

The irreversibility of infestation considerably simplifies the probability of infestation in each unit as follows:

$$P(X_{1i}|X_0) = \begin{cases} P\left(X_{1i}=0|X_{0i}=1\right) &= 0 \\ P\left(X_{1i}=1|X_{0i}=1\right) &= 1 \\ P\left(X_{1i}=1|X_{0i}=0\right) &= p_{i_{direct}} \\ P\left(X_{1i}=0|X_{0i}=0\right) &= 1-p_{i_{direct}} \end{cases}$$

Where $p_{i_{direct}}$ is the probability that i becomes directly infected from one of the units infested at time 0.

$$p_{i_{direct}} = 1 - \prod_j P_n(i,j,X_{0j},t)$$

Where X_{0j} is the initially observed status of the unit i, t is the time between the two observations and P_n is the probability of absence of infestation from unit j to unit i during time t given the infestation status x of j. Let $P_n(i,j,x,t)$ be this probability. It is calculated as:

$$P_n(i,j,x,t) = \left[1 - x \cdot p_H(i,j,t)\right]\left[1 - x \cdot p_J(i,j,t)\right]$$

Where $p_H(i,j)$ and $p_J(i,j)$ are respectively the probabilities that a hop or a jump occur during the time between the two observations from location j to location i and are related to α and ϕ by:

$$p_H(i,j,t) = I_H(i,j) \cdot \frac{\alpha \cdot t \cdot (1-\phi)}{n_H(j)}; p_J(i,j,t) = I_J(i,j) \cdot \frac{\alpha \cdot t \cdot \phi}{n_J(j)}$$

where $I_H(i,j)$ and $I_J(i,j)$ are the indicator functions for the distance between i and j to be within the hop and jump radii respectively; $n_H(j)$ and $n_J(j)$ are the number of units within hop and jump reach from j respectively; and t is the time during which the step can occur.

When considering this first order approximation in light of our data, we can see that it does not hold. Rather than many independent jump events occurring between the two observations, we instead see clusters of new invasions, suggesting that some jump steps occur which invade regions subsequently colonized by local hops.

Annexe 1.2 Intractability of higher order analytical likelihoods

Here we show how intractability arises even for the second order approximation considering that infestation can use up to one "stepping stone". That is to say that a unit can be infested by a newly infested unit only if it was directly infested by an initially infested unit.

The likelihood must then integrate over the primary infestations X_P and their times of infestations T_P to estimate the probability of the final observation:

$$L(X) = \sum_{X_P} \int_T P(X_P, T_P|X_0) P(X_1|X_P, T_P) dT_P$$

The probabilities composing this likelihood correspond to the probabilities of direct migrations and can thus be derived similarly to $P(X_1|X_0)$, but intractability arises from the integration over time for all possible primary infestation configurations.

Considering an indefinite number of dispersal steps resulting in the second observation, the definition of the likelihood function is all the more intractable as it would require to integrate over all the possible paths of infestation with any number of "stepping stones" for the infestation.

Annexe 2 Implementing the Gillespie algorithm

Given dispersal parameters α, the dispersal frequency per unit, and ϕ, the proportion of jumps, we stochastically simulate migration units occupied at the end from initially occupied units using the following Gillespie algorithm (Gillespie, 1975).

1. Draw time to next dispersal event from an exponential distribution with scale $\alpha \times n_{occupied}$, where $n_{occupied}$ is the number of occupied locations.
2. Draw the unit generating the event. We assume all occupied locations are equally likely to be the source of the migration.
3. Draw the type of event, hop or jump, according to ϕ.
4. Pick the location that will become occupied, with all locations within the hop or jump distance limits being equally likely to be selected.
5. Repeat steps 1-4 for the duration of the simulation.

We model our dispersal process as a series of independent events, and time to the next invasion in this context is given by an exponential distribution (Gillespie, 1975; Smith *et al.*, 2002). The scale parameter of the exponential distribution is $\alpha \times n_{occupied}$ because we assume that migration from each of the occupied unit occurs independently of the others.

Annexe 3 Preanalysis credible region identification

We detail our definition and computation of credible regions in the preanalysis. The 95% credible region is defined as the smallest set of points, possibly disjoint, that contain 95% of the density.

We sample the two-dimensional α, ϕ synthetic likelihood surface at regular points on a grid around the provided θ_0, the true simulation value for the dispersal parameters. At each point in the sampling grid, model simulations are run, and the density of the surface at that point is computed as described in the main text. Once the densities have been computed over the entire sampling grid, we normalize them using the overall volume underneath the computed synthetic likelihood surface, assuming that the volume is zero at points outside of the sampling grid. Then, we compute the 95% credible region as the set of points from the sampling grid with highest density and cumulatively containing 95% of the total density.

Figures A1 - A3 visually demonstrate the performance of the main sets of summary statistics considered in the preanalysis. In addition to the summary statistics presented in the main text, we additionally present the Geary's C, semivariance, annular means and variances, and different combinations of the partition based L-moment statistics.

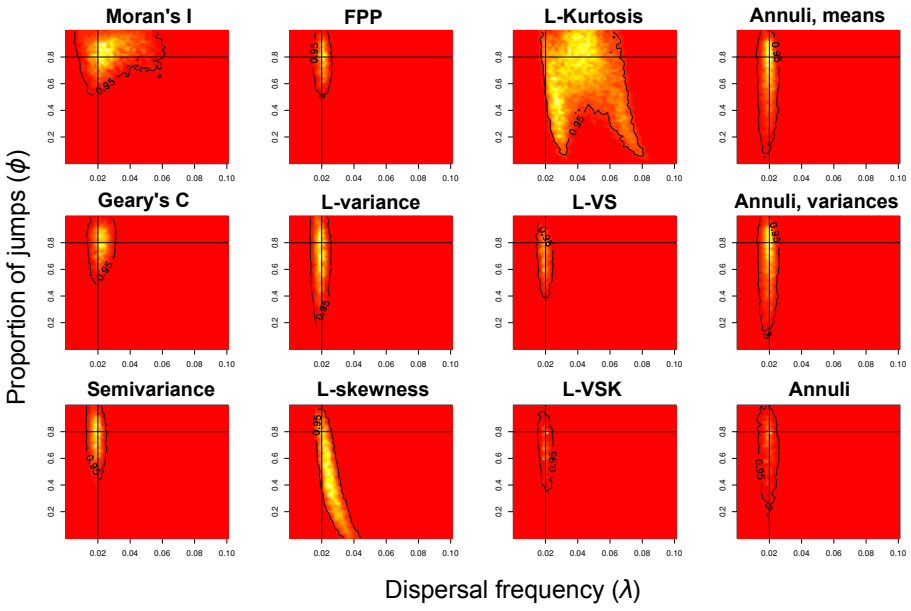

Figure A1: Preanalysis on a grid with parameters (α, ϕ): (0.2, 0.80). For each value sampled around the true value, 150 model simulations were performed. Titles for each density surface provide the summary statistic type that was used in that preanalysis.

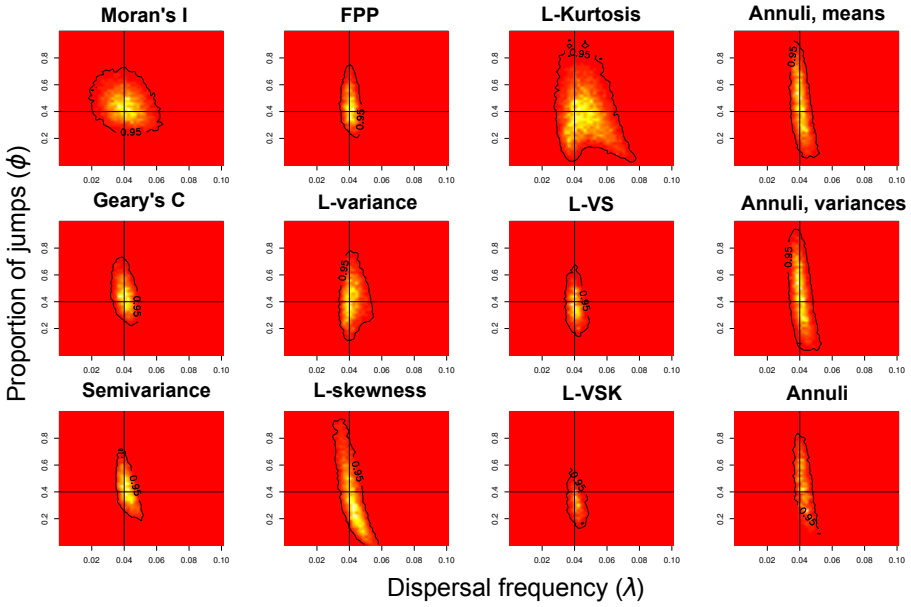

Figure A2: Preanalysis on a grid with parameters (α, ϕ): (0.4, 0.40). For each value sampled around the true value, 150 model simulations were performed. Titles for each density surface provide the summary statistic type that was used in that preanalysis.

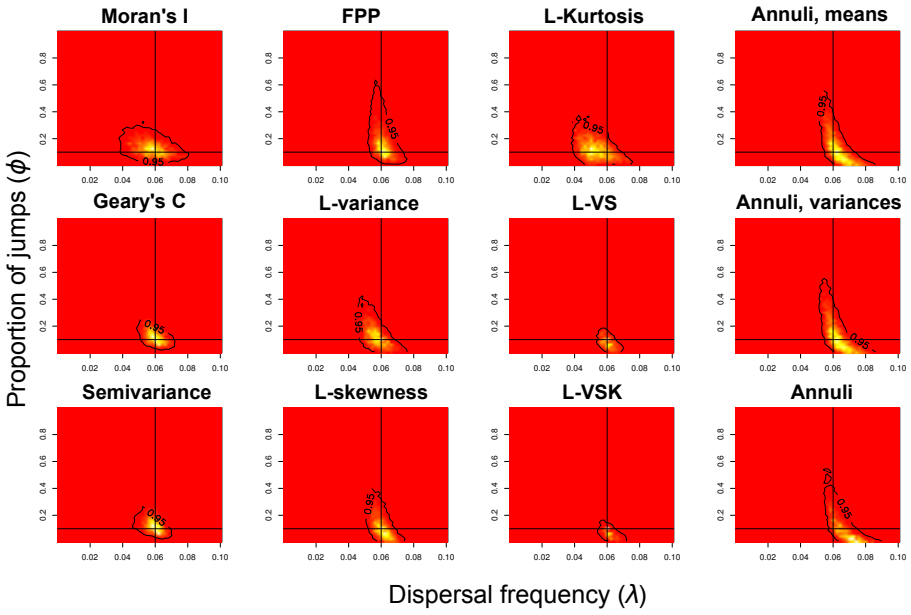

Figure A3: Preanalysis on a grid with parameters (α, ϕ): (0.6, 0.10). For each value sampled around the true value, 150 model simulations were performed. Titles for each density surface provide the summary statistic type that was used in that preanalysis.

Annexe 4 Selecting a hop jump model

As described in the main text, and for the sake of simplicity we limit our exploration of the model to two parameters, the frequency of dispersal events from a single occupied unit (α), and the probability that a dispersal event is a jump rather than a hop (ϕ). Two other parameters are nevertheless necessary to characterize the model: the hop limit, that is the distance up to which a local dispersal to neighboring units can occur, and the jump limit, the distance up to which a long distance dispersal can occur. In the case of simulations, we fix hops to be up to 30m and jumps to be between 30m and 120m. However, in our real data analysis example, we need to determine the value of these parameters.

Previous research has shown that *T. infestans* fly distances on the order of hundreds of meters when dispersing, so we allow jumping everywhere within the approximately 500m wide community (Schofield & Matthews, 1985; Vazquez-Prokopec *et al.*, 2004; Cecere *et al.*, 2006). To determine the hop limit we consider a set of possible models varying the hop limits (H_L) as 15m, 20m, ..., 60m. We parameterize each of these models using an FPP based synthetic likelihood MCMC as described in the main text. We evaluate which model, and thus which hop limit, best fits the data by estimating the Bayes Factor of these models against the best fit model (Table A1).

For every model hop limit (H_L) considered, the mean of the likelihood over the MCMC is computed. As per Kass and Raftery, the mean likelihood is used as an approximation of the likelihood of the data, D, given the model, H_L, $Pr\left(D|H_L\right)$ (Kass & Raftery, 1995). Each model is compared to the best fit model with the highest mean likelihood, H_{30} or hop limit of 30m, and the Bayes factor is computed as $BF_L = \frac{Pr(D|H_{30})}{Pr(D|H_L)}$. Bayes factors below 3 indicate support for H_L relative to H_{30}. The Bayes factor and mean likelihoods for each model and hop limit are provided in Table A1.

Hop limits of 30, 35, and 40m have the strongest support with 30m being only marginally better than 35m, and both slightly better than 40m. We find that the best H_L for the *T. infestans* data on this particular landscape is between 25 and 45m and probably closest to 30m. We thus use 30m as our hop limit for the Arequipa data and allow jumping outside of 30m.

Hop limit (m)	Mean likelihood ($\times 10^9$)	Estimated Bayes factor
15m	0.560	189
20m	1.39	76.6
25m	18.9	5.62
30m	106	1.00
35m	93.5	1.14
40m	37.7	2.82
45m	20.0	5.32
50m	3.91	27.2
55m	0.798	133
60m	0.281	375

Table A1: Estimated Bayes factors for varying hop limit models.
For every hop limit model (H_L) considered, the mean of the likelihood over the MCMC and the Bayes factor relative to the highest likelihood model H_{30} are given. Bayes factors below 3 indicate support the hop limit relative to the most likely hop limit of 30m. The set of highest Bayes factor models are 30m, 35m, and 40m, and these models provide the best parameterization of the data.

Annexe 5 Postanalysis and Preanalysis on Arequipa

After obtaining parameter estimates via synthetic likelihood MCMC for the dispersal of *T. infestans* in Arequipa, Peru, the postanalysis procedure is performed as a check of coverage and credible interval size as described in the main text. Here, the postanalysis involves simulating dispersal from initially occupied households using a provided θ_0 over the Arequipa landscape. The θ_0, the true simulation value for the postanalysis, is taken as the estimated parameter mean from the synthetic likelihood MCMC analysis. For example, when performing the postanalysis on estimates resulting from the FPP synthetic likelihood MCMC, α is taken as 0.011 and ϕ as 0.20. The specific θ_0 for the different statistics are provided in the main text and in Figure A4. Figure A4 provides the results of this postanalysis on three sets of statistics, the FPP, L-VS + N, and L-VSK + N. Coverages for all statistics considered are acceptable across both parameter sets. Additionally, the credible interval sizes for the three sets of statistics closely resemble each other.

To verify that the preanalysis provides an accurate representation of the performance of the different types of summary statistics, we compare the synthetic likelihood surface obtained from the preanalysis to that obtained from the postanalysis. As above, the θ_0, the true simulation value, is taken as the estimated parameter mean from the synthetic likelihood MCMC analysis. Postanalysis density surfaces are computed as previously detailed for preanalysis density surfaces and accompany the results of the preanalysis. The density surfaces from the preanalysis and postanalysis and 50% and 95% credible regions are provided in Figure A5.

There is a strong correspondence between the postanalysis results and the preanalysis results, respectively the upper and lower rows of Figure A5. The strong correspondence indicates that the preanalysis can confidently be used to weed out malperforming statistics before running the more computationally demanding parameter estimation and postanalysis procedures.

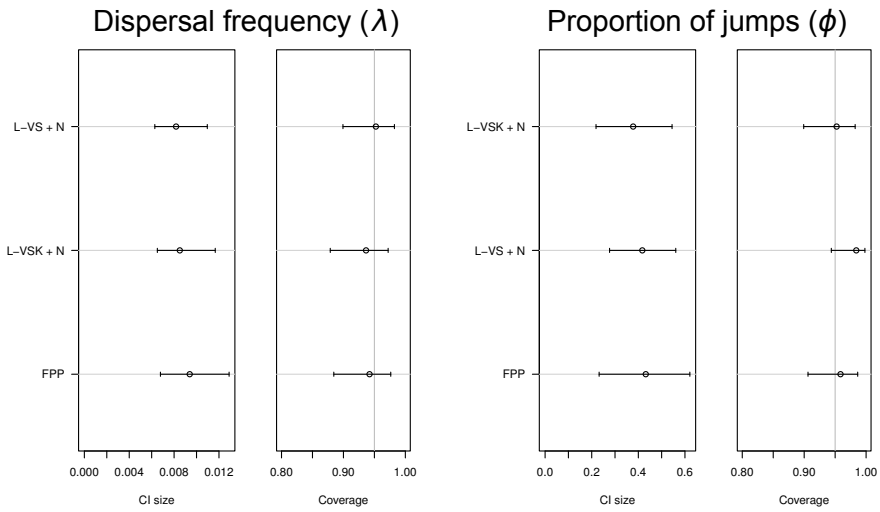

Figure A4: Credible interval sizes and coverage from postanalysis on the *T. infestans* dispersal. The postanalysis was performed with 100 MCMCs with each chain iteration consisting of 150 model simulations. θ_0 (α, ϕ), the true simulation value, is (0.011, 0.20) for the FPP statistics, (0.014, 0.30) for the L-VS + N statistics, and (0.014, 0.24) for the L-VSK + N statistics.

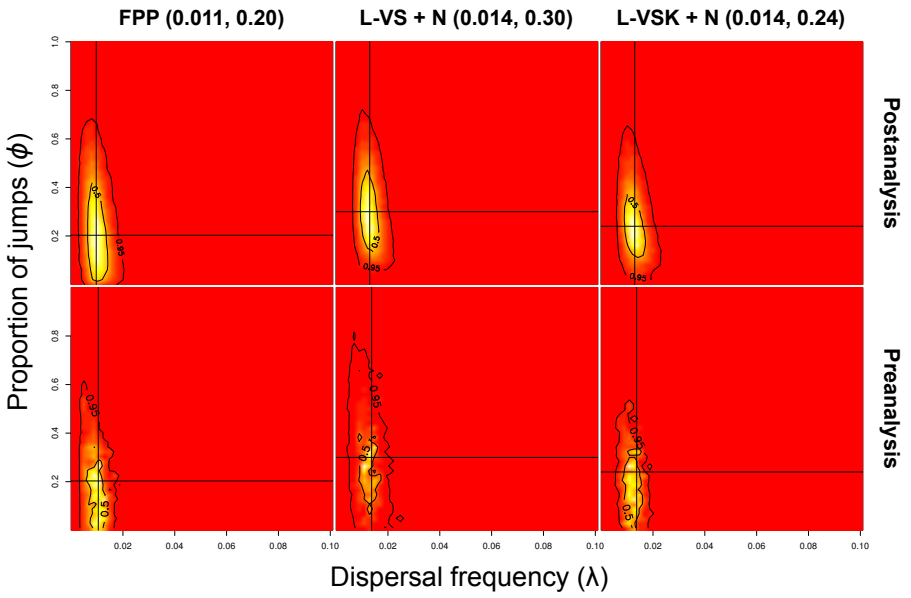

Figure A5: Correspondence between the results of the postanalysis and the preanalysis for three parameter sets. θ_0, the true simulation value, is given alongside the statistic name per column. Postanalysis results are given in the upper row and preanalysis results in the lower row. Both sets of results are given with 50% and 95% credible regions. The postanalysis was performed with 100 MCMCs with each chain iteration consisting of 150 model simulations. The preanalysis is also computed using 150 model simulations for each point analyzed around the true simulation value, θ_0.

Annexe 6 K-means partitioning

For the partition based L-moments, we divide the landscape using K-Means algorithm AS 58 Sparks (1973). This algorithm requires specification of the minimum number of units per cell n_{min}. Using the number of cells n_c and the total number of units n_u, n_min is defined as follows:

$$n_{min} = \left\lfloor \frac{n_u}{n_c} \cdot 0.85 \right\rfloor$$

Meaning that all cells have the same number of units with a tolerance of 15%.

References

1. Cecere, M.C., Vasquez-Prokopec, G.M., Gürtler, R.E. & Kitron, U. (2006). Reinfestation sources for Chagas disease vector, *Triatoma infestans*, Argentina. *Emerg. Infect. Dis.*, 12, 1096–1102.

2. Gillespie, D.T. (1975). An exact method for numerically simulating the stochastic coalescence process in a cloud. *J. Atmos. Sci.*, 32, 1977–1989.

3. Kass, R.E. & Raftery, A.E. (1995). Bayes factors. *J. Am. Stat. Assoc.*, 90, 773– 795.

4. Schofield, C.J. & Matthews, J.N. (1985). Theoretical approach to active dispersal and colonization of houses by *Triatoma infestans*. *J. Trop. Med. Hyg.*, 88, 211–222.

5. Smith, D.L., Lucey, B., Waller, L.A., Childs, J.E. & Real, L.A. (2002). Predicting the spatial dynamics of rabies epidemics on heterogeneous landscapes. *Proc. Natl. Acad. Sci. U.S.A.*, 99, 3668.

6. Sparks, D. (1973). Algorithm AS 58: Euclidean cluster analysis. *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, 22, 126–130.

7. Vazquez-Prokopec, G.M., Ceballos, L.A., Kitron, U. & Gürtler, R.E. (2004). Active dispersal of natural populations of *Triatoma infestans* (hemiptera: Reduviidae) in rural northwestern Argentina. *J. Med. Entomol.*, 41, 614–621.