Making the most of invasion records, the case of the spotted lanternfly, part II: Testing the significance of the location of jumpers, diffusers, and non-detections

Nadege Belouard[*] Sebastiona De Bona[†] Jocelyn E. Behm[‡] Matthew R. Helmus[§]

5/1/2021

Contents

[*] Temple University, nadege.belouard@temple.edu
[†] Temple University, seba.debona@temple.edu
[‡] Temple University, jebehm@temple.edu
[§] Temple University, mrhelmus@temple.edu

Aim and setup

For this second vignette, we want to test the anthropogenic character of jump events identified in the first vignette. Our working hypothesis is that SLF hitchhike on roads, railroads, and planes, and thus, establish near major transport infrastructures. To test the significance of this pattern, we measure the distance between the location of jump events (SLF hereafter called 'jumpers') and each type of transport infrastructure. Then:

- To determine whether these jumps are situated significantly closer than random to transport infrastructures, we compare the average value of this distance to a null distribution.
- To make sure that any significant result would not be due to surveys being conducted mostly close to transport infrastructures, we also calculate the distance to transport infrastructures for SLF established through diffuse spread (SLF hereafter called 'diffusers'), and for surveys that did not detect SLF (hereafter 'non-detections'). We test the significance of the distance to transport infrastructures for these two categories using random distributions again.
- Finally, we use a statistical test of means comparison to test whether jumpers are found significantly closer to transport infrastructures than diffusers or non-detections, to see if there is also a direct and significant difference between these categories.

1. Calculate distances of SLF to transport infrastructures

H1. Jump locations to transport infrastructures

We first calculate the distances of the 75 jumpers to railroads, roads, and airports.

The 75 jumpers are situated on average at 659 m of a railroad, 444 m of a road, and 6566 m of an airport.
Variables summary (min, 1st quartile, median, mean, 3rd quartile, max):
Distance to airport: 1219, 4709, 6586, 6566, 8337, 1.4351×10^4
Distance to railroad: 1219, 4709, 6586, 6566, 8337, 1.4351×10^4
Distance to road: 1219, 4709, 6586, 6566, 8337, 1.4351×10^4

H2. Diffusive spread to transport infrastructures

We select all the established populations that are not jumpers, and calculate the distance of the 5259 diffusers to railroads, roads, and airports.

The 5259 diffusers are situated on average at 3093 m of a railroad, 1198 m of a road, and 6340 m of an airport.
Variables summary (min, 1st quartile, median, mean, 3rd quartile, max):
Distance to airport: 93, 3740, 5942, 6340, 8666, 1.6524×10^4
Distance to railroad: 93, 3740, 5942, 6340, 8666, 1.6524×10^4
Distance to road: 93, 3740, 5942, 6340, 8666, 1.6524×10^4

H3. Undetected sites to transport infrastructures

We select all the points were SLF were not detected, and calculate the distance of these 35,654 points to railroads, roads, and airports.

The 35654 non-detection points are situated on average at 1.9469×10^4 m of a railroad, 1.4783×10^4 m of a road, and 2.1288×10^4 m of an airport.
Variables summary (min, 1st quartile, median, mean, 3rd quartile, max):

Distance to airport: 34, 4512, 7408, 2.1288×10^4, 1.1497×10^4, 2.489693×10^6
Distance to railroad: 34, 4512, 7408, 2.1288×10^4, 1.1497×10^4, 2.489693×10^6
Distance to road: 34, 4512, 7408, 2.1288×10^4, 1.1497×10^4, 2.489693×10^6

Hsupp. All samples to transport infrastructures

We select all the surveys, and calculate the distance of these 40,988 points to railroads, roads, and airports.

The 40,988 non-detection points are situated on average at 1.7333×10^4 m of a railroad, 1.3014×10^4 m of a road, and 1.9343×10^4 m of an airport.
Variables summary (min, 1st quartile, median, mean, 3rd quartile, max):
Distance to airport: 34, 4512, 7408, 2.1288×10^4, 1.1497×10^4, 2.489693×10^6
Distance to railroad: 34, 4512, 7408, 2.1288×10^4, 1.1497×10^4, 2.489693×10^6
Distance to road: 34, 4512, 7408, 2.1288×10^4, 1.1497×10^4, 2.489693×10^6

2. Check for differences in observed means between categories

Visual inspection

We have a first look at differences in distances to transport infrastructures between categories (Figure 1). Note that the y axis is truncated to show more clearly the boxes. Refer to the previous section to get the quartiles and maximum values of the variables. Jumpers seem to be closer to rail and roads than the other two categories, but there is no visible difference regarding the distance to airports. Diffusers are intermediate between jumpers and undetected for roads and railroads.

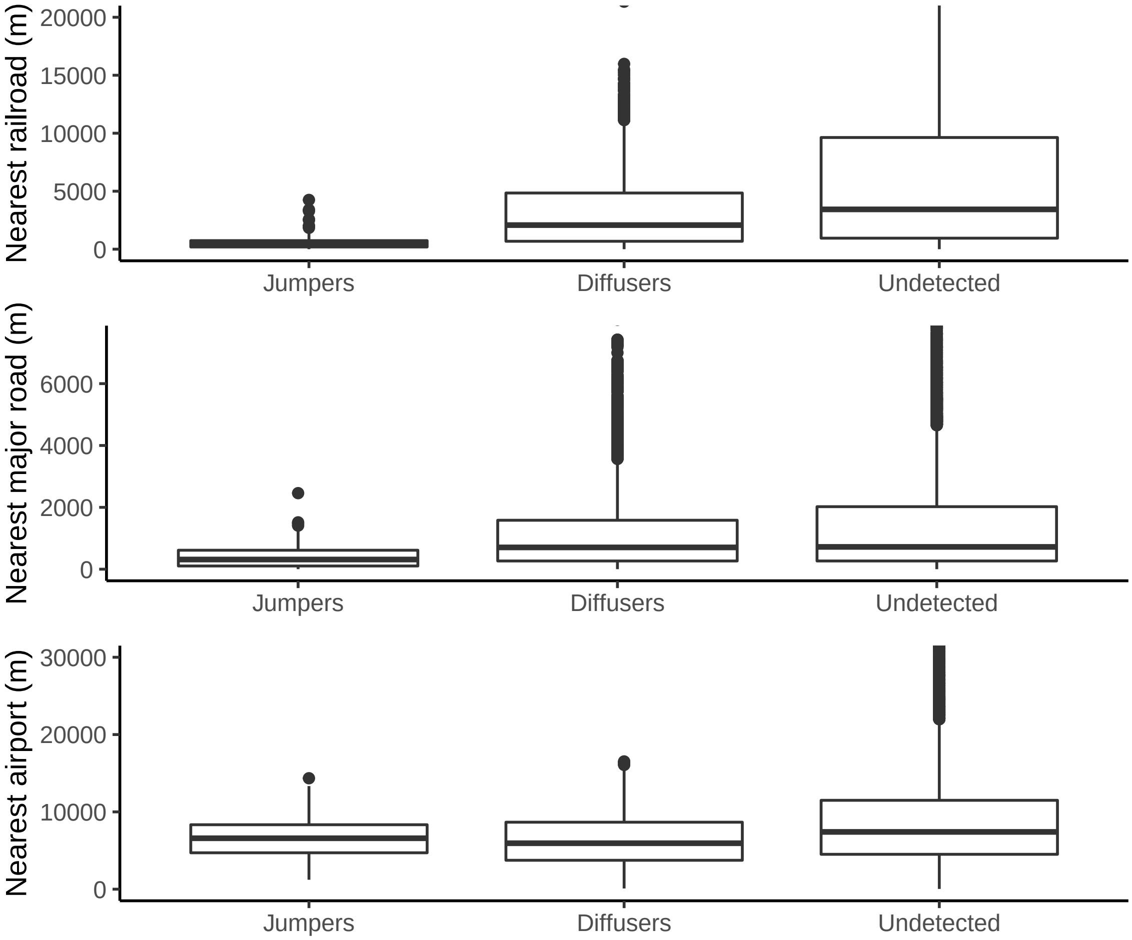

Figure 1: Distance of jumpers, diffusers and non-detections to railroads (top), roads (middle), and airports (bottom). The y axis is truncated to better show the boxes. Refer to section 1 for variables summary.

Statistical test of the difference of the means between jumpers, diffusers and non-detections

These variables are not normally distributed and the variance is not equal between groups. There is also a large difference in sample sizes (75, 5000 and 40000) Any there any statistical test that would be appropriate in this case?

Railroads

Roads

Airports

3. Generate random dispersal distribution

The idea is to generate a distribution of distances to transport infrastructures under the null hypothesis that SLF disperse randomly in the landscape. If a high number of random distributions are generated, we obtain the distribution of random dispersal distances to transports. The comparison of the random distribution and the observed average value gives the probability that the observed value is random.

Here, we generate 9,999 random datasets of distances to transport infrastructure. If the average value of the observed data is comprised within the simulated random values, it means that the pattern could be found by chance, and the distance between observed points and transport infrastructures is not significant. Clearly, it consists in: (1) generating a dataset of the same number of points as in the observed data, randomly over the same area (here, disk portions), (2) for each random point, calculating the distance to transport infrastructures, (3) taking the average distance to each transport infrastructure for the dataset, (4) redoing the same cycle again for a total of 9,999 simulated random datasets.

Note: the actual code was run on the HPC, parallelized 10 times to obtain 10,000 simulations.

H1: Random distribution of jumpers, n = 75 points

We begin with the generation of random distributions of jumpers to the three transport types: roads, railroads and airports.

We can visualize the results on Figure 2. The histogram represents the distribution of distances under the null hypothesis of random dispersal of jumpers, for each type of transport. The black vertical lines indicate the significance limits. An observed value situated outside of these vertical lines leads to the rejection of the null hypothesis. The red line indicates the average distance to transports observed in our dataset.
For all three types of transports, the observed location of jumpers is significantly closer to transports than random.

Comparison with diffusive spread and non-detections

The generation of 9,999 simulated datasets is much longer for diffusive spread and non-detections because these datasets are much bigger. As a first approach, we bootstrap the diffusers and undetected datasets, to obtain average distances based on datasets with the same number of surveys as of jumpers. We sample 75 points in the diffusers and undetected datasets (with the same share between disk portions as the jumpers), take the average distance to each transport, and repeat 9,999 times to obtain an average value that can be compared to the random distribution that we already have, before leading time-consuming simulations.

Figure 3 shows the significance of the observed distances of diffusers (blue line), non-detections (green line), and all samples confounded (yellow line) to transport infrastructures, compared to a random distribution. The results are similar for all transport types: diffusers are significantly closer to transports than random, but not as much as jumpers, and non-detections and total samples are significantly further from transports than random.

The next step would be to generate a proper random distribution for diffusers and undetected points. This is a time-consuming process, is it necessary?

Placeholder for H2: Random distribution of diffusers, n = 5,259 points

Placeholder for H3: Random distribution of non-detections, n = 35,654 points

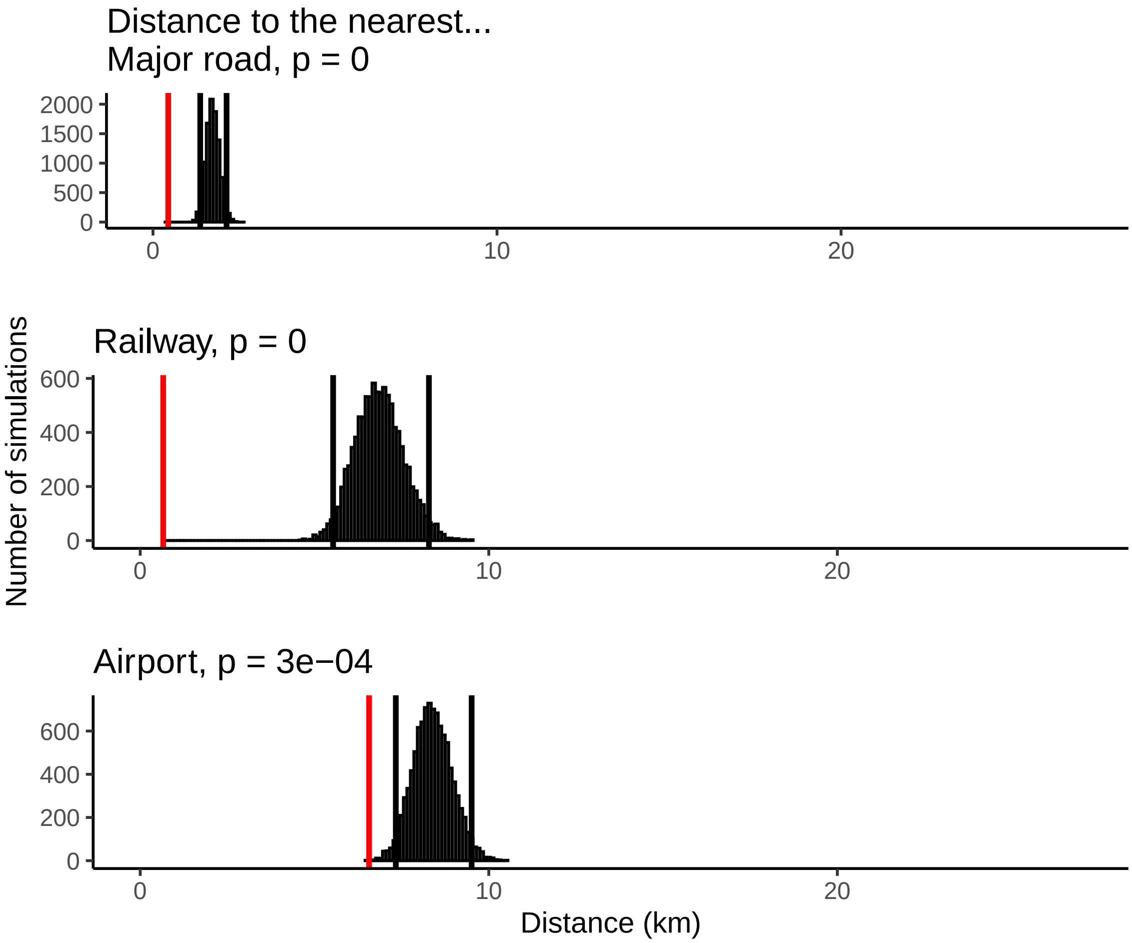

Figure 2: Comparison of the distance of jumpers to transports to a random distribution

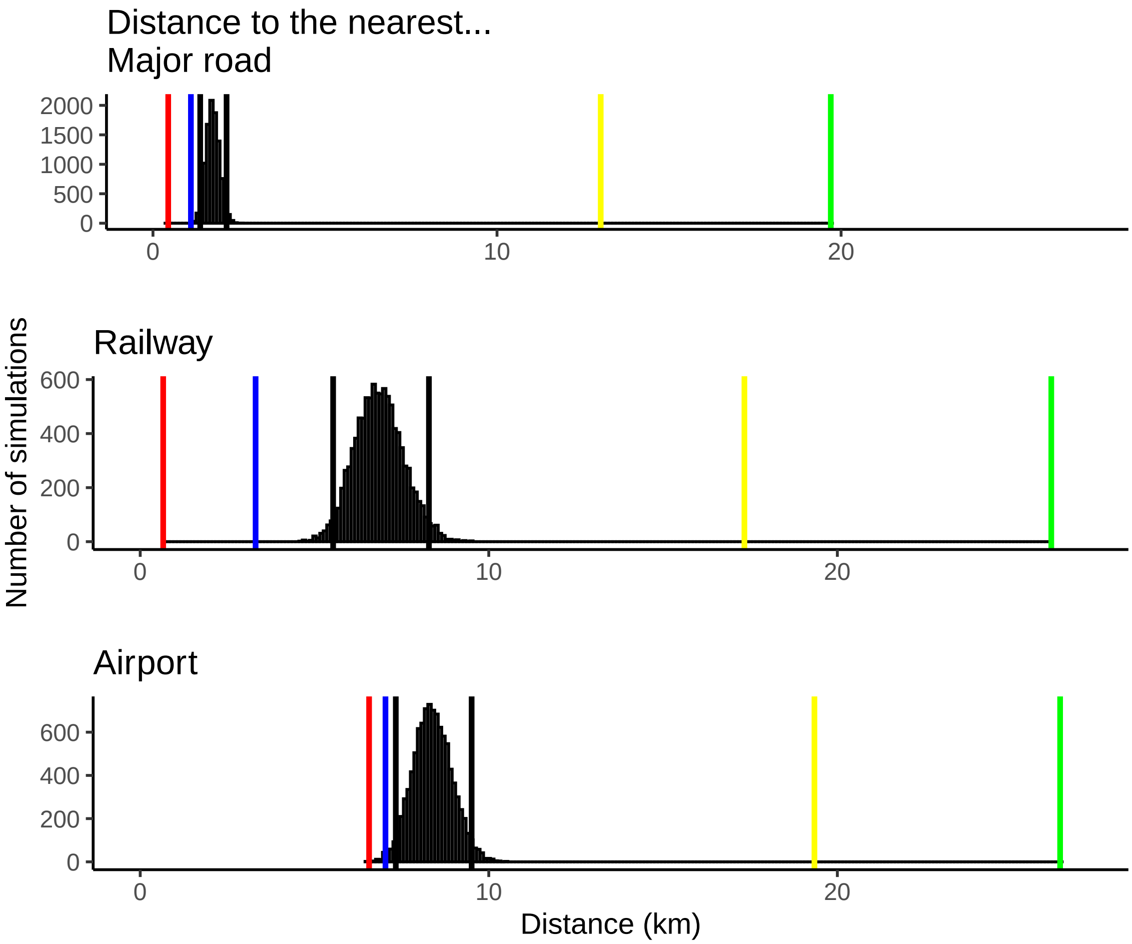

Figure 3: Comparison of the distance of jumpers, diffusers and non-detections to transports to a random distribution

4. Results and conclusion

In this vignette, we discovered that:

(0) The 40,988 surveys of the dataset that confirmed either the presence of an established SLF population or the absence of SLF have been conducted on average 13, 17 and 19 km away from major roads, railroads and airports, respectively. It is much further than random. *==> The survey method is not biased towards transport infrastructures*

(1) SLF populations, both from jump events and diffusive spread, are not located randomly, but very significantly close to transport infrastructures: roads, rail and airports. The difference to the random distribution is the highest for railroads, then roads, then airports. *==> SLF presence is tightly linked to transport infrastructures*

(2) Jump events are situated even closer to transport infrastructures than the other SLF populations, on average 444, 659 and 6,566 m away from major roads, rails and airports, respectively. *==> the establishment of new satellite populations is even more linked to transport infrastructures than diffusive spread*

(3) On the other hand, locations where SLF are not found are situated further than random from transport infrastructures. *==> SLF are unlikely to establish far from transport infrastructures*

Given these results, it is very likely that there is a causal relationship between SLF presence and transport infrastructures. SLF are likely transported by vehicles or their content, either as egg masses laid onto random materials, or at another life stage that crawls or flies on vehicles. It is likely that some transports are more involved in SLF dispersal than others. Correlation is not causation: the proximity to some transports might appear significantly related to SLF presence only because these transports are found in overall highly connected areas. This might be the case for airports, because the correlation between SLF and airports appears looser than with roads or rails. Airports are typically in areas well connected by roads and rails. More, it is assumed that adult SLF cannot survive a flight.

In the next vignette, we are going to apply this knowledge to the spatial analysis of the northeastern US, by looking at areas considered as high risk of invasion (their proximity to SLF populations and to transport infrastructures), and projecting areas likely to be invaded sooner or later.