Flying under the radar: Using spatial analytical approaches to track non-native tephritid fruit fly populations in California

By

CAROLINE C. LARSEN

DISSERTATION

Submitted in partial satisfaction of the requirements for the degree of

DOCTOR OF PHILOSOPHY

in

Ecology

in the

OFFICE OF GRADUATE STUDIES

of the

UNIVERSITY OF CALIFORNIA

DAVIS

Approved:

James R. Carey, Chair

Robert J. Hijmans

Jay a. Rosenheim

Committee in Charge

2021

Acknowledgements

One accumulates a great deal of both academic and personal knowledge over the course of earning a graduate degree.

The academic. The joke that graduate school is the process of learning more and more about less and less was certainly true in my experience, though it disregards the positive side of this: there is great joy and privilege in exploring specialized topics of particular interest and mastering their skills. The following three chapters highlight a sliver of those that I honed over the course of my doctoral experience, though certainly not the entire breadth.

The personal. The doctoral journey is on paper (both literally and figuratively) a purely academic pursuit: the personal journey is far more difficult, far more transformative, and far more rewarding. There's immense room for self-discovery while undertaking such a massive task as a dissertation, but the simple passage of time means that significant personal life events are likely to occur. In times of emotional trauma, it becomes impossible to separate one's personal journey from their professional one. In my times of trauma and joy while in graduate school, I have been lucky to be held by both my personal community and my academic community. There are innumerable people without whom I would not have been able to complete my doctoral degree, much less begin it in the first place. Acknowledgements cannot do justice to the immense gratitude I have for everyone who has touched my life in both monumental and small ways.

First, I would like to thank my advisor, James Carey, for his encouragement and trust over the last ten years. I appreciate the balance of academic space and structure he provided me, as well as his support of my independent development of research projects, his prompt and constructive feedback, his mentorship, and his patience. I could neither have started nor completed this journey without him. I would also like to thank my co-advisor, Robert Hijmans, for unlocking my interest in geospatial ecology. Without his course and personal guidance, I would not have been able to meld my inherent interest in geospatial methodology with my dataset. I thank him for his enthusiasm during times I felt overwhelmed and his hours of coding help in R which made wrestling my difficult dataset feel more like fun problem-solving opportunities.

Many other people dedicated their time and energy to supporting my research and academic journey as well. I would like to thank the rest of my dissertation committee and my qualifying exam committee: Jay Rosenheim, Scott Carroll, Phil Ward, Marissa Baskett, and Peter Moyle. I appreciate their time, feedback, guidance, and their enthusiasm for their diverse fields of expertise. And, of course, for making my qualifying exam experience an exciting, supportive environmental rather than a terrifying one. I appreciate their time, thoughtful feedback, and ability to broaden my scientific perspective. Thank you to many others at U.C. Davis who positively impacted my time here: Holly Hatfield, Pat Randolph, Sharon Lawler, and Neal Williams. Finally, thank you to those ecologists who first inspired me to pursue ecological research during my undergraduate career: Laurence Kruger and Karen Vickers. In the middle of the South African savanna, they guided me through a dark place and back to the happiest version of myself, filled with adventure and surrounded by bugs.

I am forever indebted to the boundless friendship and immeasurable support of Ash Zemenick, Jenny VanWyk, Alex Webster, Allison Simler, and Sacha Heath during the hardest of times. Their selfless sacrifices during a time of multiple group traumas kept me alive. I love and cherish each of their beautiful hearts and brilliant minds. Along with these five, I would like to thank the Ecology Graduate Group and related programs which served as my academic home for nearly a decade. So many incredible souls and minds made my time here special. Particular thanks to Kate Tiedeman, Brendan Barret, Emily Rothwell, Julia Michaels, Lily Tomkovic, Hannah Waterhouse, and Billy Krimmel for gifting my life with their exuberance.

Finally, I offer thanks to my family, without whom I would not be here at all. To my artist mother, Malinda, for fostering my love of bugs, science, and exploration, despite her severe aversion to butterflies. To my engineer father, Charles, for always cheering me on, even when I stubbornly refused to answer, "How's the dissertation coming?" To my sister, Austen, for her levity, cheese plates, editing prowess, and for being my oldest friend. To Ash, for trauma bounding, friendship traps, and pizza. To Otto, for teaching me to howl at the moon. To Michael, who I miss dearly: without him I would never have begun this journey. And to Rebecca, my love, my wife: without her I could never have finished it. I am so excited for our future together.

None of this work would have been possible without funding and institutional support. Thank you to the National Science Foundation Graduate Research Fellowship Program, the Henry A. Jastro Graduate Research Scholarship, the U.C. Davis Graduate Group in Ecology, and the U.C. Davis Department of Entomology and Nematology.

Abstract

Despite intensive and organized efforts to track species of high importance to humans, population censuses are often unable to detect small extant populations. For threatened or endangered species, this may result in conservative estimates of populations that bolster conservation efforts. However, for pests or non-native and potentially invasive species, failing to properly consider small, early-stage invasion populations can result in foregoing intervention strategies until populations are already established. Research on difficult-to-detect insect pests is critical to understanding the population dynamics of potentially harmful species before the negative effects of their full impact are realized. To address this latter issue, in this dissertation I utilize a highly unique dataset that has tracked non-native fruit fly (Diptera: Tephritidae) populations in California for over one hundred years to explore spatial statistical techniques that aid in detecting populations that are typically sub-detectable. In Chapter 1, I use spatial point pattern analysis with a variety of temporal treatments to confirm potential establishment signals of *Bactrocera dorsalis* (the oriental fruit fly) in the Los Angeles region of California. Building on this, Chapter 2 uses point pattern clustering metrics to determine which non-native tephritid species are likely already established in California and therefore pose a higher risk of invasion or outbreak. Finally, in Chapter 3, I explore which human and environmental factors drive *B. dorsalis* detections in California. Combined, these analyses reveal deeper dynamics of difficult-to-detect populations of non-native tephritid fruit flies and provide an approach for monitoring early-stage invasive species. Looking forward, these findings may inform best management practices as global climate change and globalization increase the likelihood of further problematic insect introductions worldwide.

Table of Contents

Introduction

Invasive species research is critical to understanding the population dynamics of potentially harmful species before the negative effects of their full impact are realized. The word "harmful" is subjective: invasive species, regardless of origin, have been associated with numerous negative ecological effects (e.g., biodiversity loss, changes in ecosystems services) and negative economic effects (e.g., forestry, fisheries, and agriculture).[1–4] Invasive and potentially-invasive species posing economic risk are often the subjects of large economic investment for research, monitoring, and management. A prime example of this are non-native tephritid fruit flies. Heavy monitoring efforts for more than a century have generated a large spatiotemporal dataset of non-native tephritid detections in California, but there has been little analysis of this dataset for deeper ecological understanding or predictive applications. In the following three chapters, I reveal deeper dynamics of non-native tephritid populations in California and provide methodological approaches for analyzing limited population data over time.

Background and Motivation

Tephritid fruit flies (Diptera: Tephritidae), also called peacock flies or true fruit flies, are a diverse group encompassing over 4,600+ currently-described species globally.[5,6] The majority of tephritid species, especially non fruit-eating species, exist peacefully in their native ranges. However, about 70 frugivorous species are considered serious threats to agriculture worldwide outside of their home ranges. In fruit-eating species, females oviposit under the skin of fruit and the larvae consume the fruit flesh until pupation, thus ruining the plant for human consumption. These species are heavily monitored, managed, and regulated throughout trading channels.[5]

In a worst-case scenario, each of the high-risk tephritid species not only has the potential to destroy entire sections of the agricultural industry but can prevent salvageable portions of crops from being exportable due to strict international trade regulations, especially with countries that do not yet have those potentially troublesome species. The strict policy surrounding tephritids means that the mere presence of some of these species would be enough to cause massive quarantines and to close some export markets entirely.[5,7] As of 2020, California's economy is ranked approximately 5th largest in the world. The state's agricultural sector is a major component: California is the world's 5th largest supplier of "food and agricultural commodities."[8] Though California is also home to many benign native tephritid species, the CDFA considers non-native tephritids the "most important agricultural pest in the world."[5,9]

Despite the somewhat ominous nature of insect invasions, every ecological change is an opportunity for unplanned study of the dynamics of species in novel environments. The spread of non-native tephritids in California is no exception. The necessary monitoring of these species has resulted in a long-term dataset that can be used for ecological analysis with broad implications for better understanding invasion dynamics. Over time, *invasive* has become a loaded term that carries the deterministic assumption that once colonizers of particular species arrive in a suitable area, a full-force invasion is imminent, leaving only eradication or a doomed ecosystem as potential outcomes.[2,10] The inclusion of small, early-invasion populations in the invasion ecology paradigm is important because it implies that many if not most non-native species exist undetected for long periods of time. Little is known about small populations in the context of invasion ecology. By virtue of being frequently sub-detectable, they are difficult to study using traditional methods. Further, research that does exist is heavily skewed towards species of

conservation interest. Very few non-native species explode in numbers immediately after their introduction.[2,11–13] This lag gives non-native species time to naturalize and form positive as well as negative community interactions.[12,14,15] This also has major implications for how we view eradication. Eradication is generally an immediate response to detection of an undesirable species. However, if species can exist unseen for long period of time before detection (and they often do) it is likely that a population would have spread beyond the treatment boundary by the time treatment is deemed necessary. Thus, from a management perspective, true eradication is extremely difficult and rare despite common declaration that insect pests have been eradicated locally and regionally.[2,16,17] In reality, many species categorized as highly invasive may lurk below the detection threshold for a long period of time, similar to what we believe is the case with non-native tephritid species in California.

Finally, studying the long-term dynamics of tephritid populations can provide insights into how urban and suburban development impact invasion. Tephritids are an agricultural pest, but California agriculture is kept tephritid-free through intensive preventative management. The populations of concern in this dissertation are persisting in developed areas of the state where human-supported backyard fruit trees and vegetable gardens increase the potential niche of the species. By the year 2050, more than 70% of the global population is expected to live in urban zones.[18] In the United States, 80% of people already live in urban and suburban areas.[19] Development is detrimental to some species, though many may be able to adapt. Increasing human populations and the suburban sprawl consuming the state is a benefit to non-native tephritids. The larger the matrix, the more able pests are to find survival reservoirs, more opportunities for refuge from post-detection control measures. This in turn increases the likelihood of outbreaks in close proximity to agricultural areas.

Approach

The dataset. The tephritid detection dataset that serves as the backbone of this dissertation is unusually rare in its biological, temporal, and spatial depth. Due to California's agro-economic importance, the U.S. Department of Agriculture (USDA) and California Department of Food & Agriculture (CDFA) spend more than $20M per year on non-native tephritid monitoring and detection.[9,20] The product is a network of 100,000+ traps across California checked weekly for the last 75 years. This enormous sampling effort (over $1B in total) is generating a high-resolution dataset containing the date, exact coordinates, and sex of each individual non-native tephritid captured across the state (n=11,000+).[21–24] Previous analysis suggests as many as nine of the 17 monitored non-native tephritid species are established and slowly spreading across California, yet are persisting at densities so low they are usually sub-detectable outside of occasional small-scale outbreaks that are immediately suppressed,[25–27] though not without debate.[28,29] These studies proved that although each individually detected fly is a rare event, it may nevertheless represent a diffuse, but established, population. One can imagine each detection or outbreak as an island breaching the ocean surface. The challenge becomes how to infer the topography of the rest of the submerged mountain range far below those few peaks, i.e., hidden below detectable densities.

A dataset so fine-scale and long-term such as this one is a rare and powerful tool in ecology that has several unique attributes. First, the number of species in the dataset is exceptional. To date, 17 species across only three genera of tephritid fruit flies have been found in the state. Exhaustive lists of invasive insects show that species are often the sole invasive representatives of their genus or even families. The coincidence of so many non-native tephritid species in

California alone, across four genera, is therefore highly intriguing. The dataset includes the species, sex, and life stage of each individual fly, providing opportunity to compare detection patterns across multiple species and genera.

Second, the data spans a century in time. Long term ecological datasets are hard to come by, especially for introduced species. Whereas most historical occurrence data must be pieced together from sources such as museum collections and citizen science reporting, the first-hand tephritid data stretches back decades, predating the arrival of many tephritid species. This temporal depth is unusual in its coverage of the entire invasion timeline and thereby provides an opportunity to study early invasion stages. In contrast, the invasion literature focuses predominately on the later phases of invasion (rapid growth, spread, and negative impact).

Finally, the spatial depth and breadth of the data is astounding. The baited traps making up the trapping network are placed at densities of approximately 5 traps per square mile of developed, non-agricultural regions of the state. This amounts to roughly 100,000+ traps across the state checked weekly from spring to fall. The location of each individual adult or larvae is documented with precise coordinates. For context, imagine finding a single fly in a football field. Now, imagine finding a single fly in a state the size of nearly 80 million football fields across, each year for a century. Even given the power of baited or pheromonal lures, the statistical power of each individual detection is staggering.

Despite its many advantages, there are also characteristics of the dataset that make analysis difficult. The presence-only data, the variable sampling effort, the effects of post-detection control, and the small sample size warrant careful consideration in data analysis. First, the data is presence-only: there is no information about traps that did *not* receive a detection, only those that did. This presented unique challenges for analysis and interpretation that are discussed in each chapter. Second, trapping procedures have been constant since the 1980s, but earlier information is difficult to find. In most cases, I assume that trapping effort, in terms of trap density and collection frequency, has remained stable over time, but I discuss where this assumption is not made or otherwise impacted analyses. Third, post-detection control measures may have partially eradicated or dampened populations, but explicitly measuring or modeling this effect was beyond the scope of this work. Instead, I consider this effect one of many possible external factors influencing tephritid populations by modeling the detections themselves, which are a sample of the underlying population. Finally, though the dataset is generated from a trapping array that is large and high-resolution, and the full dataset constitutes over 11,000 individual flies, the annual abundance for each individual species is quite small. Small sample sizes are a ubiquitous statistical problem as they are problematic for many tests and may not cover enough variation to fully represent the underlying population.[30–32] Most approaches for modelling rare species emerged from species of conservation concern. Conversely, invasion models are often designed for rapid population growth or spread.[33–37] The tephritid data represents a small sample of a slow-spreading, difficult-to-detect population. We have modeled the data in each chapter to account for the small sample size by binning data, using moving windows, or using statistical methods that do not rely on large samples sizes.

Three Studies. Model-based approaches are increasingly relied upon in applied ecology, and I aim to add to the body of knowledge surrounding non-native tephritids in California using a variety of models that complement the unique characteristics of the dataset.

Chapters 1 and 2 rely on the principles of point pattern analysis. Point pattern analysis broadly concerns relationships among a set of *events* (here, detections) in a defined study area (broadly, California).[38–41] In general, point patterns can be statistically abstract, but in *spatial point pattern analysis* these events consist of geographic locations with or without associated information (*marks*). A frequent goal of point pattern analysis is to determine the degree of clustering (aggregation) in an observed point pattern. By incorporating a temporal dimension, we use these techniques to explore difficult-to-detect populations of a non-native tropical fruit fly species.

In Chapter 1, we examine the *Bactrocera dorsalis* detection data for space-time clustering, an indication of potential establishment. We take a simple approach by using *nearest neighbor analysis* (NNA), which considers the distance from each event to its closest neighboring event.[38,39] Small observed distances represent clustering, while larger or more even distributions of distances suggest randomness or regularity in the observed pattern. We test the importance of temporal grouping in determining the degree of space-time clustering, an indication of early-stage establishment. We further examine whether early-stage clustering signals differ significantly from random distributions.

In Chapter 2, we hypothesize that despite having limited detection data, occasionally-detected species can be still be monitored for spatiotemporal aggregation, an early sign of establishment, over time. We posit that these occasionally-detected species will exhibit clustering characteristics intermediate to both rarely- and frequently- detected species. We use spatial statistics to differentiate which, if any, species merit being considered higher risk of establishment based on clustering metrics. Spatial ecology suggests that different kinds of spatial statistics are best used in concert due to their varying strengths, sensitivities, and limitations.[42,39,43] We use two separate spatial statistics, the *L* function and the O-ring statistic, to analyze the spatiotemporal patterns of six non-native tephritid species: frequently-detected *Bactrocera dorsalis* and *Ceratitis capitata* and occasionally-detected *Anastrepha ludens*, *Bactrocera correcta*, *Bactrocera zonata*, and *Bactrocera cucurbitae*.

Chapter 3 moves away from point pattern analysis into more traditional linear modeling with the goal of statistical inference. We examine the distribution of *Bactrocera dorsalis* fruit flies in the Los Angeles, CA area. Using a combination of natural, human, and *B. dorsalis* population variables, we identified the strongest explanatory variables of detections using random forest and logistic regression. These variables lend support to our belief that *B. dorsalis* is established in the Los Angeles region over alternative hypotheses of new introductions.

In this research, I aim to understand the underlying spatiotemporal dynamics of non-native tephritids using limited occurrence data from detection trapping grids. Examining both the large- and small-scale spread patterns of various species will provide insight into invasion dynamics

but will also explore how we can infer deeper dynamics from limited, secondary data. Though many elements of the tephritid data in California are unique, there are also ubiquitous principles that represent an overlap between invasion ecology, agricultural ecology, urban ecology, modeling rare species. This work would not be possible without an interdisciplinary approach, and I encourage this to extend into the management realm of these species.

Works cited

1. Szyniszewska, A. M. & Tatem, A. J. Global assessment of seasonal potential distribution of Mediterranean fruit fly, Ceratitis capitata (Diptera: Tephritidae). *PLoS One* **9**, (2014).
2. Davis, M. A. *Invasion Biology*. (Oxford University Press, 2009).
3. *Encyclopedia of Biological Invasions*. (University of California Press, 2011).
4. Ehrenfeld, J. G. Ecosystem Consequences of Biological Invasions. *Annu. Rev. Ecol. Evol. Syst.* **41**, 59–80 (2010).
5. Triplehorn, C. A., Johnson, N. F. & Borror, D. J. *Borror and DeLong's Introduction to the Study of Insects*. **7**, (Thomson Brooks/Cole, 2005).
6. Marshall, S. A. *Flies: The Natural History & Diversity of Diptera*. (Firefly, 2012).
7. Lance, D. R., Woods, W. M. & Stefan, M. Invasive Insects in Plant Biosecurity: Case Study – Mediterranean Fruit Fly. in *The Handbook of Plant Biosecurity* 447–484 (2014).
8. U.S. Department of Agriculture. *California Agricultural Statistics*. (2012).
9. *Medfly infestation triggers quarantine in central Los Angeles*. (2014).
10. Davis, M. A. *et al.* Don't judge species on their origins. *Nature* **474**, 153–154 (2011).
11. Hastings, A. *et al.* The spatial spread of invasions: new developments in theory and evidence. *Ecol. Lett.* **8**, 91–101 (2004).
12. Crooks, J. Lag times and exotic species : The ecology and management of biological invasions in slow-motion. *Ecoscience* **12**, 316–329 (2005).
13. Simberloff, D. The Role of Propagule Pressure in Biological Invasions. *Annu. Rev. Ecol. Evol. Syst.* **40**, 81–102 (2009).
14. Templeton, A. R. The reality and importance of founder speciation in evolution. *BioEssays* **30**, 470–9 (2008).
15. Davis, M. A. *Invasion biology*. (Oxford University Press, 2009).

16. Liebhold, A. M. & Tobin, P. C. Population ecology of insect invasions and their management. *Annu. Rev. Entomol.* **53**, 387–408 (2008).

17. Carroll, S. P. Conciliation biology: the eco-evolutionary management of permanently invaded biotic systems. *Evol. Appl.* **4**, 184–199 (2011).

18. Loss, S. R., Ruiz, M. O. & Brawn, J. D. Relationships between avian diversity, neighborhood age, income, and environmental characteristics of an urban landscape. *Biol. Conserv.* **142**, 2578–2585 (2009).

19. U.S. Census Bureau. *Decennial Census Summary File 1*. (2000).

20. Gilbert, A. J., Bingham, R. R., Nicolas, M. A. & Clark, R. A. *Insect Trapping Guide*. (2013).

21. Papadopoulos, N. T., Plant, R. E. & Carey, J. R. From trickle to flood : the large-scale , cryptic invasion of California by tropical fruit flies. *Proc. R. Soc.* (2013).

22. Carey, J. R. Establishment of the Mediterranean Fruit Fly in California. *Science (80-.).* **253**, 1369–1373 (1991).

23. Carey, J. R. The future of the Mediterranean fruit fly Ceratitis capitata invasion of California: A predictive framework. *Biol. Conserv.* **78**, 35–50 (1996).

24. Chen, I. From Medfly to Moth: Raising a Buzz of Dissent. *Science (80-.).* (2010).

25. Carey, J. R. The Mediterranean Fruit Fly. Invasion of California Deepens: Response to an alternate Explanation for Recurring Outbreaks. *American Entomologist* (2010).

26. Papadopoulos, N. T., Plant, R. E. & Carey, J. R. From trickle to flood: the large-scale, cryptic invasion of California by tropical fruit flies. *Proc. Biol. Sci.* **280**, 20131466 (2013).

27. Zhao, Z. *et al.* Life table invasion models: spatial progression and species-specific partitioning. *Ecology* **100**, 1–11 (2019).

28. Gutierrez, A. P., Ponti, L. & Gilioli, G. Comments on the concept of ultra-low , cryptic tropical fruit fly populations. *Proc. R. Soc. B* **281**, (2014).

29. Carey, J. R., Plant, R. E. & Papadopoulos, N. T. Response to commentary by Gutierrez et al . Response to commentary by. 8–10 (2014). doi:10.1098/rspb.2013.2825.

30. Mi, C., Huettmann, F., Guo, Y., Han, X. & Wen, L. Why choose Random Forest to predict rare species distribution with few samples in large undersampled areas? Three Asian crane species models provide supporting evidence. *PeerJ* (2017). doi:10.7717/peerj.2849

31. Pearson, R. G., Raxworthy, C. J., Nakamura, M. & Townsend Peterson, A. Predicting species distributions from small numbers of occurrence records: A test case using cryptic geckos in Madagascar. *J. Biogeogr.* **34**, 102–117 (2007).

32. Cunningham, R. & Lindenmayer, D. Modeling count data of rare species: some statistical issues. *Ecology* **86**, 1135–1142 (2005).

33. Sakai, A. *et al.* The population biology of invasive species. *Annu. Rev. Ecol. Syst.* **32**,

305–332 (2001).

34. Neubert, M. G. & Parker, I. M. Projecting rates of spread for invasive species. *Risk Anal.* **24**, 817–31 (2004).

35. Hastings, A. *et al.* The spatial spread of invasions: new developments in theory and evidence. *Ecol. Lett.* **8**, 91–101 (2005).

36. Lewis, M. A. Invasion Biology. in *Encyclopedia of Theoretical Ecology* (eds. Hastings, A. & Gross, L.) 365–391 (University of California Press, 2012).

37. Hastings, A. Models of spatial spread: A synthesis. *Biol. Conserv.* **78**, 143–148 (1996).

38. Diggle, P. J. *Statistical analysis of spatial and spatio-temporal point patterns, third edition*. (Taylor & Francis Group, LLC, 2013). doi:10.1201/b15326

39. O'Sullivan, D. & Unwin, D. J. *Geographic Information Analysis: Second Edition*. (John Wiley & Sons, Inc, 2010). doi:10.1002/9780470549094

40. Renner, I. W. *et al.* Point process models for presence-only analysis. *Methods Ecol. Evol.* **6**, 366–379 (2015).

41. Ripley, B. D. Modelling Spatial Patterns. *J. R. Stat. Soc. Ser. B* **39**, 257–267 (1977).

42. Diggle, P. J. On Parameter Estimation and Goodness-of-Fit Testing for Spatial Point Patterns. *Biometrics* **35**, 87–101 (1979).

43. Perry, G. L. W., Miller, B. P. & Enright, N. J. A comparison of methods for the statistical analysis of spatial point patterns in plant ecology. *Plant Ecol.* **187**, 59–82 (2006).

How time flies: Point pattern analysis reveals temporal persistence of *Bactrocera dorsalis* populations in California

Caroline C. Larsen-Bircher, Robert J. Hijmans, James R. Carey

Abstract

Oriental fruit flies (Tephritidae: *Bactrocera dorsalis*) are a global pest and major focal species in invasion biology research. In this study, we add to previous research demonstrating the establishment of this agricultural pest species in California using point pattern analysis. We test the impact of various temporal groupings on clustering metrics of *B. dorsalis* detection patterns. By adapting classic point pattern analysis tools to incorporate a temporal dimension, we (1) confirm the importance of comparing detection patterns at multiple time scales to determine presence of difficult-to-detect populations and (2) provide further evidence of oriental fruit fly establishment in the state.

Introduction

As biological invasions increase in frequency and severity in step with global climate change, [1–5] invasion research will become more critical to understand the dynamics of potentially damaging species. California accounts for over two thirds of the fruit and nuts and over one third of the vegetables produced in the US. The state has consequently invested heavily in the prevention, monitoring, and management of invasive agricultural pests.[6] Non-native tephritid fruit flies

account for a large proportion of those invaders, and *Bactrocera dorsalis* has proven to be one of the most dominant tephritid species.[7–10] *Bactrocera dorsalis*, commonly known as the oriental fruit fly, has expanded from its native range in Asia to become one of the most globally widespread pest species.[8,11,12] *B. dorsalis* owes its growing geographic range, and therefore its extensive economic impact, to its highly polyphagous behavior (130+ known host crops), climatic adaptability, and impressive dispersal ability.[11–15] The first *B. dorsalis* detection in California occurred in 1960 in the city of Anaheim over half a century ago, and the species has been detected annually in the state since 1969. Cumulative annual detections vary significantly, yet 75% of years from 1970 to 2014 experienced at least five detections. According to the California Department of Food and Agriculture (CDFA), the combined value of potential agricultural hosts at risk to *B. dorsalis* would be over $16.4 billion (as of 2015).[12]

Extensive state and federal resources are devoted to prevention, monitoring, and management of agricultural pests.[16,17] However, mounting evidence suggests that many non-native tephritid species, including *B. dorsalis*, are established in small pockets in the state of California. Papadopolous *et al.* (2013) compared tephritid interception rates at international and domestic ports of entry as a measure of propagule pressure to locations of past outbreaks and detections of various tephritid species. The authors generated a county- and local-scale recapture model to test the hypothesis of establishment. They determined that between five and nine non-native tephritid species may be established in California. Zhao *et al.* (2019) examined tephritid populations in California using life table invasion models and found the number of infested cities to be steadily increasing. They further determined that invasion outcomes depend on which species is first detected in a given area. These studies proved that although each individually detected fly is a

rare event, it may nevertheless represent a diffuse, but established, population. This means that each data point holds great statistical power. Our research further considered the statistical power of each individual detection in the context of point pattern analysis.

Spatial and spatiotemporal analysis are critical tools in invasion biology: determining the risk of a problematic species necessarily involves examining biogeographic elements of the population over time.[18–21] Myriad techniques have been developed to best understand historical, current, and forecasted population dynamics, ranging from simple distance-based-measures to complex forecasting. However, many methods are not well-suited for analysis of long-term datasets. Most point pattern statistics are designed for events within a single time period or between events in two distinct time periods. Species with relatively infrequent detections over long time periods require a different approach. In this paper, we will explore difficult-to-detect populations of a non-native tropical fruit fly species using simple-but-effective nearest neighbor analysis, an underused tool for consideration of early-stage and difficult-to-detect invasions. We ask whether increasing the temporal window of analysis improves our ability to detect signals of clustering, and therefore possible establishment. We further examine whether early-stage clustering signals differ significantly from random distributions.

Methods

Database & Study Species

Tephritid fruit fly populations have been heavily monitored in California since their first detection in Hawaii over a century ago (*Bactrocera cucurbitae* in 1895; *Ceratitis capitata* in 1907).[8,11] Since then, a trapping network of both pheromone and baited traps has formed a fine-

scale detection grid across most of the state.[16,17] We used the corresponding historical dataset provided by the CDFA and supplemented by public records and reports from the Plant Health & Pest Prevention Services branch. Each detection record in the database represents an individual fly found in California over the past century and includes the species, capture date (day, month, year), and geographic coordinates (latitude, longitude). This date ranges from 1960 through 2014. Though this version of the dataset ends in 2014, tephritids have been found in California annually since then as well. We focused on *B. dorsalis* due to its detection frequency over the last 60 years. The following analysis, modeling, and mapping procedures were implemented in R programming language.[22,23]

In this paper, we use three variations of this database. The first is the raw spatial data, including every individual detection event, which we refer to as the *full dataset*. We also use an *outbreak-collapsed dataset,* where any detection events with identical coordinates within the same year are collapsed into a single event. Since the CDFA moves the pest traps multiple times throughout the active season, same-trap detections indicate an outbreak in a short time period. Collapsing these outbreaks into single events minimizes skewing in distance-based measures due to over-representation of specific locations in the dataset. Finally, we also generated a *random coordinates dataset* by randomly drawing coordinate pairs from within the local-scale study area (defined below) using the R package *sp*.[24,25] We matched the number of random events per year to the number of observed events in the outbreak-collapsed dataset per year (within the study region) to minimize the effect of variable yearly abundance on distance metrics.

Study Area

We define two study areas in this research. We refer to the full range of every *B. dorsalis* detection in the database, the entire state of California, as the *regional-scale* study.[26] This area has defined (political) boundaries, but analyses conducted at this larger scale are not sensitive to edge effects since the entire dataset is represented.[20] The *local-scale* study area is a smaller polygon in Los Angeles. We selected this area for two reasons. First, in order to test observed data patterns against a null hypothesis of random introductions, a carefully defined study area is necessary.[20,27,28] Second, since the regional boundaries of cumulative *B. dorsalis* presence is extremely irregular, largely due to topography, we chose a small study area within a region of Los Angeles. General trapping densities are stable within human-dominated areas of the state – fewer traps exist in less populated regions.[16] Defining a study area within the large, continuous sprawl of Los Angeles helps ensure the detections are the result of a similar trapping effort. Further, Los Angeles has both many of the oldest and most recent detections, creating a large temporal range.[14]

Point Pattern Analysis

Point pattern analysis broadly concerns relationships among a set of *events* in a defined study area. While general point patterns can be statistically abstract, in *spatial point pattern analysis* these events consist of geographic locations with or without associated information (*marks*). A frequent goal of point pattern analysis is to determine the degree of clustering (aggregation) in an observed point pattern. This is done by comparing a given point pattern to a null model, often *complete spatial randomness* (CSR). Statistically, a point pattern is the observed realization of an underlying stochastic process that can be defined by its *first order* and *second order properties*. First order properties refer to the variable intensity of the underlying process across space.

Second order properties, on the other hand, reflect the relationships among the events themselves, once first order variation is accounted for.

Regional-Scale Nearest Neighbor Analysis

One pillar of point pattern analysis is *nearest neighbor analysis* (NNA). NNA evaluates the spatial distance between a focal event and the event closest to it, its eponymous *nearest neighbor.*[19,20] Small observed distances represent clustering, while larger or more even distributions of distances suggest randomness or regularity in the observed pattern. While basic in concept, we find nearest neighbor analysis to be an intuitive and adaptable tool for exploring rare event spatial clustering. Further, while nearest neighbor analyses (and point pattern analyses more broadly) are most frequently conducted within a small temporal period, we demonstrate its potential as a tool for detecting space-time clustering in long-term datasets in the context of non-native *B. dorsalis* in the state of California.

In this research, we use the term *focal year* (FY) to define the primary set of events (individual tephritid detections) being analyzed. In *same-year* comparisons, this FY is compared to itself: i.e., an NND is found for each detection in the given year to any other detection occurring within that same year. In *inter-year* comparisons, a FY is compared to a defined set of *prior years* (PY), which range between the one year and the five years prior to the FY.

We first explored the influence of temporal groupings in NNA of the entire outbreak-collapsed dataset for the regional-scale study area. Same-year comparisons were calculated using a symmetric distance matrix of the detection pattern from each FY against itself using the R

package *raster*.[29] We then conducted a series of inter-year comparisons using an asymmetric distance matrix to compare the detection pattern of a given FY to the detection pattern of a cumulative PY range. For all same- and inter-year comparisons, the NND was calculated for every detection event in the full dataset.

We evaluated these results in two ways. First, we calculated the minimum NND and the median NND for each year and each comparison type to determine how these values changed over time. For example, in the same-year comparison, spatial coordinates of all detections in focal year 1990 were compared to all non-self detections in that same year. In the 3 years prior comparison, the distribution in 1990 was compared to the distribution of all detections from years 1987 to 1989. The focal year ticked forward to 1991 and the process repeated. Second, we considered the effect of temporal distance on NNDs. We took a full distance matrix of all detection events and determined the smallest nearest neighbor distance event in each year to every other individual year. We plotted these minimum NNDs as a function of temporal distance between each year being compared.

Local-Scale Nearest Neighbor Analysis

A benefit of nearest neighbor analysis is that it can avoid the necessity of a defined study area: focusing on smallest point-to-point distances in an entire observed point pattern minimizes the underestimation of neighbor density or skewing of inter-point distances due to first order process variation across a larger area.[19–21] However, this makes testing against a null distribution difficult. To provide an example of a more localized, hypothesis-based spatial test, we used our local-scale study area boundary, the outbreak-condensed dataset (subset to study area), and the random coordinate dataset. We repeated the same- and inter-year comparison framework as in

the regional-scale analysis with the subset observed data, and again with the random data. For example, in the same-year comparison, spatial coordinates of all detections in focal year 1990 were compared to a randomly generated distribution with the identical number of detections. In the three years prior to comparison, the distribution in 1990 was compared to random distributions with the same number of detections as observed from 1987 to 1989. The focal year ticked forward to 1991 and the process repeated. We calculated the minimum NND and the median NND per year for each and plotted the results in the same figure windows.

Chi-Squared Analysis

We tested whether past distributions influence future distributions, and whether detections with near neighbors in prior years were more likely to have near neighbors in future years, compared to a new detection (i.e., detection with no prior near neighbors). For each individual detection in the full dataset, we calculated the spatial distance from that detection to all other detections in the preceding five years and then for the following five years within two "spatial thresholds" (S=5km and S=15km). Detections falling within the same year as the focal detection were not included. We binned counts of the number of neighboring detections in the temporal range and spatial thresholds into categories of 0 (none), 1-5 (some), and 6+ (many) neighbors. We then compared these previous and future near neighbor counts using a Chi-Squared test.

Results

The raw detection data were mapped in R using WGS 84 coordinates.[29–31] The first set of maps (Figure 1.1a) divides all detections in California by decade. The second set depicts cumulative detection distribution in the greater San Francisco Bay and Sacramento areas (Figure 1.1b) and in the greater Los Angeles and San Diego areas (Figure 1.1c). The maps show first two decades

of detections only in southern California, slowly increasing in range. *B. dorsalis* spreads to northern CA in the 1980s and remains there.

We plotted cumulative *B. dorsalis* detections over time in both the complete dataset and the outbreak collapsed dataset (Figure 1.2). We found a total of 1319 individual detections in the complete dataset from 1960 to 2014. The outbreak-collapsed dataset constituted 976 separate detection events. Apart from a large outbreak in 1974, annual cumulative detections increase from the 1960s to 1980s, then remain fairly stable with occasional outbreaks. Differences between the two datasets are minimal but indicate outbreaks occurring at some scale every few years. The majority of the detections are captured from individual traps.

Regional-Scale Nearest Neighbor Analysis

Nearest neighbor distributions of *B. dorsalis* in California vary over time and depend on both the number of detections in a given year (high detection years are statistically more likely to have a close neighbor) and the specific distributions being compared. Detection patterns cluster more strongly (i.e., have smaller NNDs) within that same year than they do with the prior year (Figure 3a). However, when we compare a given year to a longer prior range, we see the range of NNDs decrease substantially, indicating substantial clustering in space and time. This signals that a detection in a given year may be more likely to have a closer neighbor from two to five years before than the year immediately prior. Localized populations likely grow to a detectable level with some evenness at very small scales, followed by a depressed detection year due to post-detection control. Figure 4 demonstrates that while the very smallest minimum NNDs (minimum NND < 0.1 km) occur from same-year comparisons (temporal distance=0), there is

little other effect of temporal distance on minimum NND: detections can have close neighbors regardless of temporal distance between detection patterns.

Local-Scale Nearest Neighbor Analysis

In this analysis, we used events from the outbreak-condensed dataset within the defined study area (Figure 1.5), as well as randomly generated data as described above. For each of the four comparison categories (same-year, PY=1, PY=3, and PY=5), minimum (Figure 1.6) and median (Figure 1.7) NNDs observed-to-random comparisons were larger (i.e., less tightly clustered) than those in observed-to-observed comparisons. This demonstrates that observed detection patterns of *B. dorsalis* within the study area are more highly clustered both within and between years that would be expected under the null hypothesis of random detection. As the random dataset was generated using the same number of detections per year as the spatially-subset outbreak-collapsed dataset, the major spikes due to abnormally high or low detection years are similar.

Chi-Squared Analysis

The results of the smaller 5km spatial threshold indicate that new detections (detections with no neighbors in the preceding five years) were equally likely to continue to have no neighbors (48%) as they were to have some neighbors (1-5 = 49%; 6+ =3%) in the future (Table 1.1a). Detections with some (1-5) prior neighbors are more than twice as likely to continue to have at least one neighbor in the future (68%) than they are to have none (32%). Detections with a high number of previous neighbors are most likely to have some neighbors (56%) in the future but are equally likely to have either no neighbors (24%) or many neighbors (19%) in the future.

At the larger 15km spatial threshold, we find that new detections are similarly likely to have no future neighbors as with the 5km threshold (42%; Table 1.1b). However, out of new detections

which have future neighbors, there is a far higher proportion of new detections having many future neighbors as opposed to some (1-5 = 35%; 6+ = 23%) compared to the 5km threshold. 94% of detections that had many (6+) future neighbors at this threshold had a minimum of one near neighbor in the past.

At the 5km threshold, only 25% of detections have both zero previous and zero future neighbors. The percent in this category decreases substantially, to 7%, at the 15km scale. In contrast, the proportion of detections with many past neighbors and many future neighbors is only 2% of the total in the 5km threshold, this increases to 31% at the 15km threshold. At either scale, detections are far more likely to have a similar number of, or more, future neighbors (80% at 5km; 87% at 15km) than they are to have fewer future neighbors (19% at 5km; 13% at 15km).

Discussion

This study reveals the importance of how time factors into analysis of invasion dynamics. From a management perspective, the success of control measures is determined by whether there are repeat occurrences in the year following individual detections or outbreaks. However, this study indicates that comparing patterns in adjoining years may be misleading for management. Weak clustering signals between two adjoining years, especially following an outbreak, may reflect a temporary effect of localized control measures rather than the eradication of the entire population (Figure 1.3; Figure 1.7).[8,10,16] This is reflected in our results: patterns in adjoining years show little clustering, while cumulative patterns of the three or five years preceding a given year generate very strong clustering signals. Having very close spatial neighbors separated by two to

five years in time has multiple implications. First, it suggests that post-detection control may not be effective long-term, despite short-term efficacy at depressing small surging populations. Second, consideration of multiple years after a detection or outbreak can reveal space-time clustering and potential establishment signals that may be unseen with short-term post-detection monitoring. These points combined highlight the danger of considering detections and small outbreaks "eradicated" after only a year or two of monitoring.

The above analyses further confirm signals of likely establishment of *B. dorsalis* in the state of California through the following three lines of evidence: (1) regional-scale clustering, (2) local-scale clustering, and (3) Chi-squared analysis. Detection patterns within a given year generally show high levels of clustering. This follows naturally as individual flies resulting from the same low-abundance population would likely be found in close proximity to each other, as opposed to being more randomly distributed within the range as would be expected under the null hypothesis of independent introductions (Figure 1.3). We see further signals of establishment from our local-scale analysis. *B. dorsalis* detection patterns are continually more tightly clustered both within and between years than would be expected under a null random distribution, another strong indication of subtle but ever-present establishment signal (Figure 1.6; Figure 1.7).

New detections at a small spatial scale seem to have a relatively equal chance of non-recurrence (48% at 5km; 42% at 15km) and having future neighbors (52% at 5km; 58% at 15km), which particularly at the 5km scale could indicate an established population (Table 1.1). This ratio presumably increases as the spatial threshold is widened. Detections that have many neighbors

within 5km are likely experiencing an outbreak in the focal year. These outbreaks events would experience subsequent outbreak mitigation and eradication efforts, which in turn influences the chance of future neighbors in the short term. At either scale, it is far more likely for any given detection to have the same amount of or more neighbors in the future than it is for a detection to have fewer neighbors in the future. Some of these patterns increase naturally as the spatial threshold widens: detections are more likely to have more neighbors in a wider range than in a smaller range. However, since this applies to both prior and future detections, the percentages in each category are still worth exploring. An individual *B. dorsalis* can easily fly 10km and they have been known to travel as far as 30km.[12,32]

The high stakes of *B. dorsalis* invasion in California from its (a) economic risk due to broad host range (230+ crops known)[11,12] and (b) short, rapid life cycle means/implies/suggests that a comprehensive understanding of *B. dorsalis* populations is critical. Interpretation of data many times over has suggested that *B. dorsalis* is established in California: though the analytical approaches vary, results of each study clearly indicate levels of establishment in urban and suburban areas.[8,10] This research has shown that (a) in most areas, once they arrive, they then stay; (b) while annual detection counts on average aren't dramatically increasing, the spatial area both in terms of spread and in-fill continues to increase; and (c) there appear to be some predictive trends. Our research adds to this by confirming signals of establishment based on fine-scale inter-point distance analysis. Future research will aim to further quantify specifics of where these pockets of establishment are in space and time, and predictive covariates of established and non-established space.

Yet there remains considerable disagreement on tephritid establishment in California – one reason among many why this research is critical.[33–36] Detections per year are not increasing rapidly (Figure 1.1)[8,10] and are stable at levels far below peak annual abundance. Studies indicate the efficacy of preventative measures in import and exports.[8,37,38] This is an indication that despite *B. dorsalis*'s patchy establishment, populations remain successfully suppressed. Nevertheless, acknowledgement of tephritid establishment in California could result in catastrophic embargo on agricultural exports. This friction between ecology and politics highlights the inefficiency of all-or-nothing pest status. Ecologically, the best management approach with such a high-risk species would err on the side of caution and consider the worst-case scenario (in this case, establishment). Holistic consideration of these populations will strengthen our long-term management, and therefore economic, success. In future studies, we will expand this point pattern methodology to other tephritid species in order to determine comparative risk.

Acknowledgements

We would like to thank Dr. Jay Rosenheim, Dr. Ash Zemenick, and Dr. Jenny VanWyk for helpful feedback on the manuscript and Austen Wright for editing. We thank the Carey and Hijmans lab for advice. This material is based on work supported by National Science Foundation Graduate Research Fellowship Program Grant. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation. Specific author contributions: CCL conceived the study, analyzed data, and wrote the initial draft. RJH assisted with coding

and analysis. JRC provided the dataset, expertise on the study species, and contributed substantial feedback at each stage of the project.

Works Cited

1. Walther, G.-R. *et al.* Alien species in a warmer world: risks and opportunities. *Trends Ecol. Evol.* **24**, 686–93 (2009).
2. Robinet, C. & Roques, A. Direct impacts of recent climate warming on insect populations. *Integr. Zool.* **5**, 132–42 (2010).
3. Biber-Freudenberger, L., Ziemacki, J., Tonnang, H. E. Z. & Borgemeister, C. Future risks of pest species under changing climatic conditions. *PLoS One* **11**, 1–17 (2016).
4. Dukes, J. S. *et al.* Responses of insect pests, pathogens, and invasive plant species to climate change in the forests of northeastern North America: What can we predict?This article is one of a selection of papers from NE Forests 2100: A Synthesis of Climate Change Impacts o. *Can. J. For. Res.* **39**, 231–248 (2009).
5. Kokko, H. & López-Sepulcre, A. From individual dispersal to species ranges: perspectives for a changing world. *Science* **313**, 789–91 (2006).
6. California Department of Food & Agriculture. California Agricultural Statistics Review 2017-2018. (2018).
7. Bateman, M. The ecology of fruit flies. *Annu. Rev. Entomol.* **17**, (1972).
8. Papadopoulos, N. T., Plant, R. E. & Carey, J. R. From trickle to flood : the large-scale , cryptic invasion of California by tropical fruit flies. *Proc. R. Soc.* (2013).
9. Aluja, M. & Norrbom, A. *Fruit Flies (Tephritidae): Phylogeny and Evolution of Behavior*. (CRC Press, 2010).
10. Zhao, Z. *et al.* Life table invasion models: spatial progression and species-specific partitioning. *Ecology* **100**, 1–11 (2019).
11. Stephens, A. E. A., Kriticos, D. J. & Leriche, A. The current and future potential geographical distribution of the oriental fruit fly, Bactrocera dorsalis (Diptera: Tephritidae). *Bull. Entomol. Res.* **97**, 369–378 (2007).
12. California Department of Food & Agriculture. Oriental Fruit Fly Fact Sheet. (2018). Available at: https://www.cdfa.ca.gov/plant/factsheets/OFF_FactSheet.pdf.
13. Froerer, K. M. *et al.* Long-Distance Movement of Bactrocera dorsalis (Diptera: Tephritidae) in Puna, Hawaii: How far can they go? *Am. Entomol.* **56**, 88–94 (2010).
14. Papadopoulos, N. T., Plant, R. E. & Carey, J. R. From trickle to flood: the large-scale, cryptic invasion of California by tropical fruit flies. *Proc. Biol. Sci.* **280**, 20131466 (2013).

15. Carey, J. R., Papadopoulos, N. T. & Plant, R. Tephritid pest populations oriental fruit fly Outbreaks in California: 48 consecutive years, 235 Cities, 1,500 detections-and counting. *Am. Entomol.* **63**, 232–236 (2017).

16. Gilbert, A. J., Bingham, R. R., Nicolas, M. A. & Clark, R. A. *Insect Trapping Guide*. (2013).

17. California Department of Food & Agriculture & U.S. Department of Agriculture. Exotic Fruit Fly Regulatory Resonse Manual. (2001). doi:10.1093/infdis/jit776

18. *Conceptual ecology and invasion biology: reciprocal approaches to nature*. (Springer Netherlands, 2006). doi:10.1007/1-4020-4925-0

19. Diggle, P. J. *Statistical analysis of spatial and spatio-temporal point patterns, third edition*. (Taylor & Francis Group, LLC, 2013). doi:10.1201/b15326

20. O'Sullivan, D. & Unwin, D. J. *Geographic Information Analysis: Second Edition*. (John Wiley & Sons, Inc, 2010). doi:10.1002/9780470549094

21. Velázquez, E., Martínez, I., Getzin, S., Moloney, K. A. & Wiegand, T. An evaluation of the state of spatial point pattern analysis in ecology. *Ecography (Cop.).* **39**, 1042–1055 (2016).

22. R Core Team. R: A language and environment for statistical computing. (2013).

23. RStudio Team. RStudio: Integrated Development for R. (2016).

24. Bivand, R. S., Pebesma, E. & Gómez-Rubio, V. *Applied Spatial Data Analysis with R*. (Springer New York, 2013). doi:10.1007/978-1-4614-7618-4

25. Pebesma, E. & Bivand, R. Classes and methods for spatial data in R. *R News* **5**, (2005).

26. Gaston, K. J. *The Structure and Dynamics of Geographic Ranges*. *Oxford Series in Ecology and Evolution* (2003). doi:10.2167/jost191b.0

27. Gaston, K. J., Jackson, S. F., Cantú-Salazar, L. & Cruz-Piñón, G. The Ecological Performance of Protected Areas. *Annu. Rev. Ecol. Evol. Syst.* **39**, 93–113 (2008).

28. Diggle, P. J. *Statistical analysis of spatial and spatio-temporal point patterns, third edition*. (Taylor & Francis Group, LLC., 2013). doi:10.1201/b15326

29. Hijmans, R. J. raster: Geographic Data Analysis and Modeling. (2019).

30. Wickham, H. *ggplot2: Elegant Graphics for Data Analysis*. (Springer-Verlag New York, 2016).

31. Kahle, D. & Wickham, H. ggmap: Spatial Visualization with ggplot2. *R J.* **5**, 144–161

32. Froerer, K. M. *et al.* Long-distance movement of Bactrocera dorsalis (Diptera: Tephritidae) in Puna, Hawaii: How far can they go? *Am. Entomol.* **56**, 88–95 (2010).

33. Carey, J. R., Papadopoulos, N. & Plant, R. The 30-Year Debate on a Multi-Billion-Dollar Threat: Tephritid Fruit Fly Establishment in California. *Am. Entomol.* (2017).

34. Carey, J. R. The Mediterranean Fruit Fly. Invasion of California Deepens: Response to an alternate Explanation for Recurring Outbreaks. *American Entomologist* (2010).

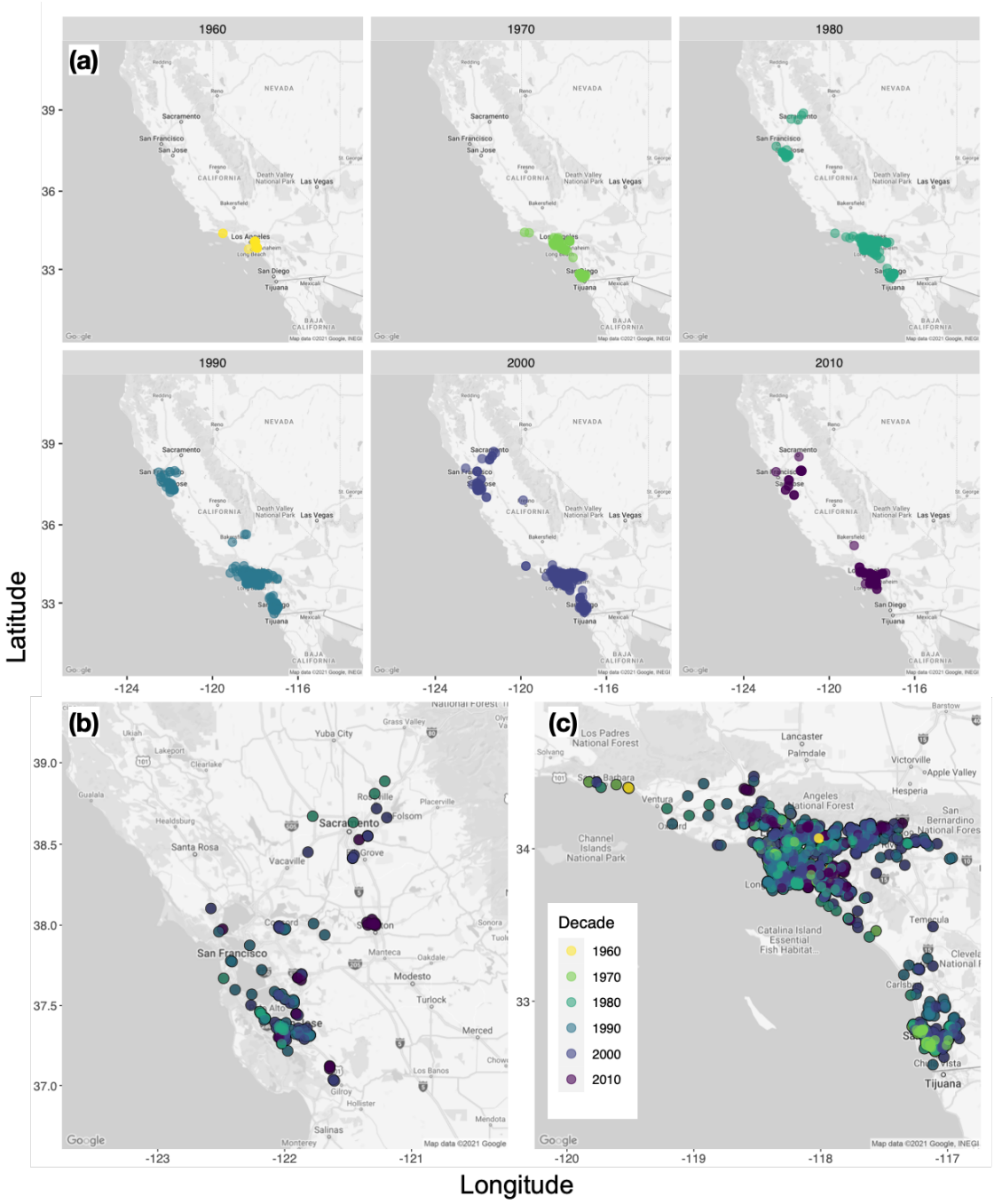
35. McInnis, D. *et al.* Can polyphagous invasive tephritid pest populations escape detection for years under favorable climatic and host conditions? *Am. Entomol.* **63**, (2017).

36. Shelly, T. E. *et al.* To Repeat: Can Polyphagous Invasive Tephritid Pest Populations Remain Undetected For Years Under Favorable Climatic and Host Conditions? *Am. Entomol.* **63**, 224–231 (2017).

37. Tobin, P. C. *et al.* Determinants of successful arthropod eradication programs. *Biol. Invasions* **16**, 401–414 (2013).

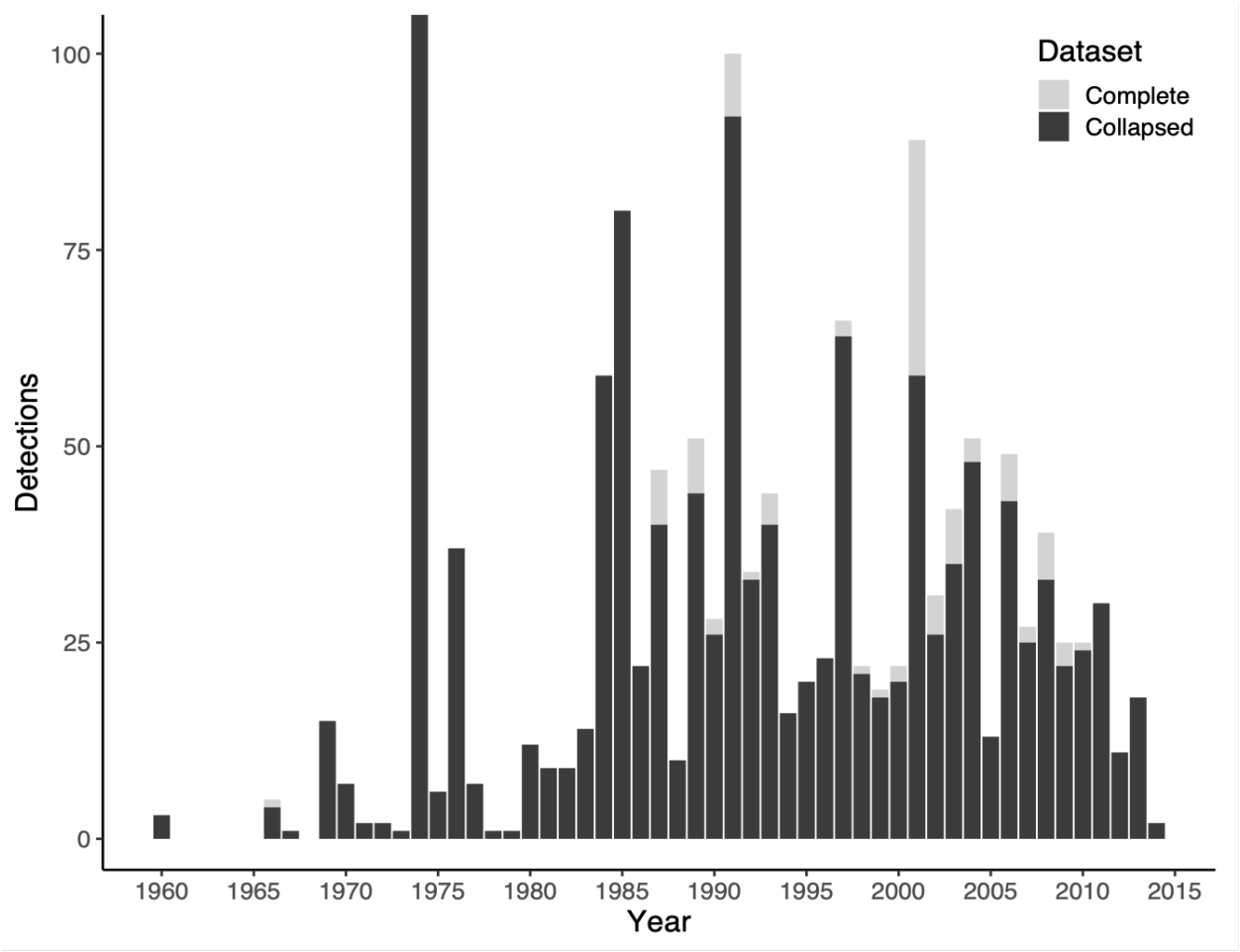
38. Suckling, D. M. *et al.* Eradication of tephritid fruit fly pest populations: outcomes and prospects. *Pest Manag. Sci.* (2014). doi:10.1002/ps.3905

Figure 1.1.



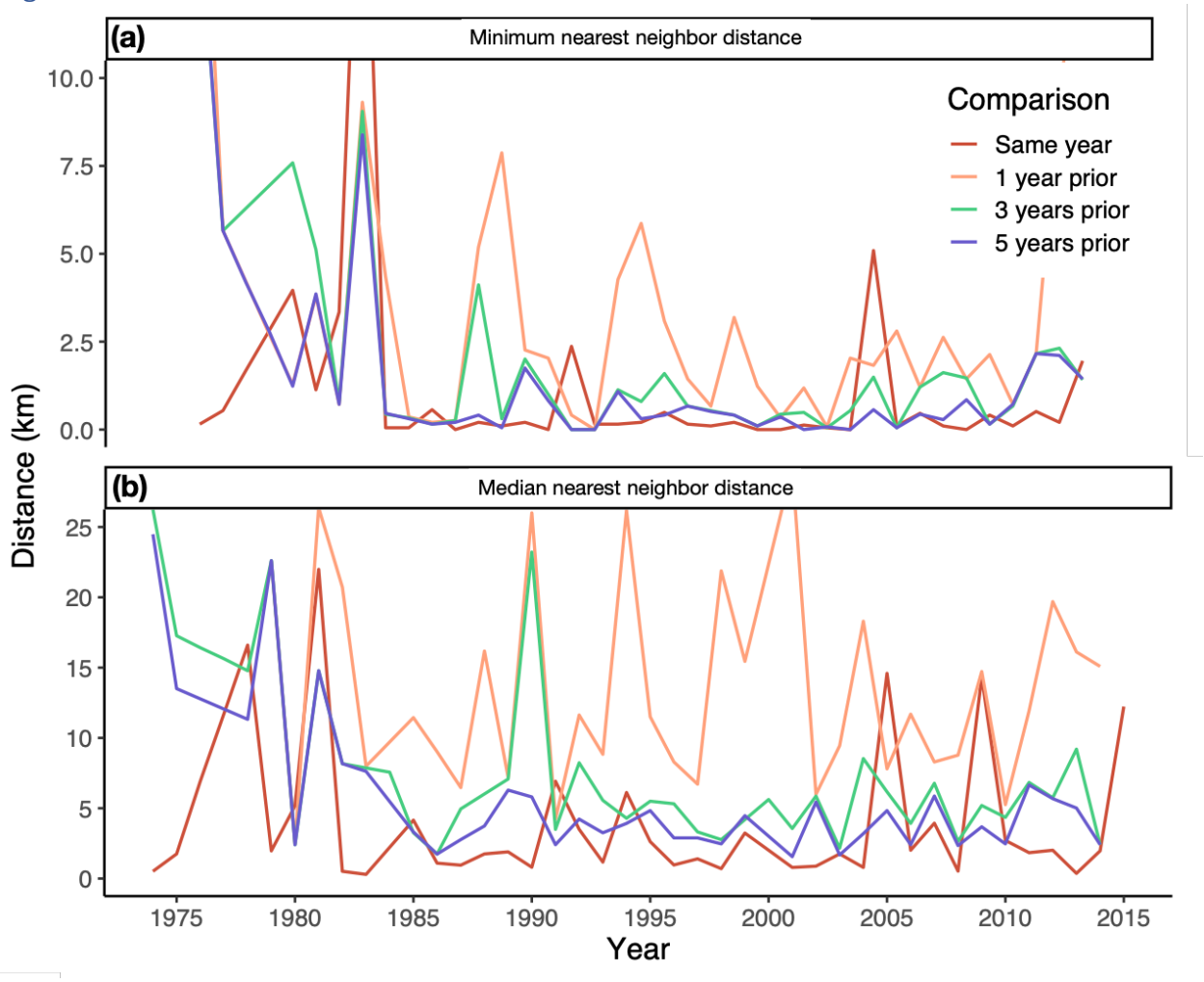
Summary of B. dorsalis detections in California, with data color-coded by decade. Maps depicting all B. dorsalis detections in (a) all California by decade; (b) the greater San Francisco Bay and Sacramento areas cumulatively; and (c) the greater Los Angeles and San Diego areas cumulatively.

Figure 1.2.



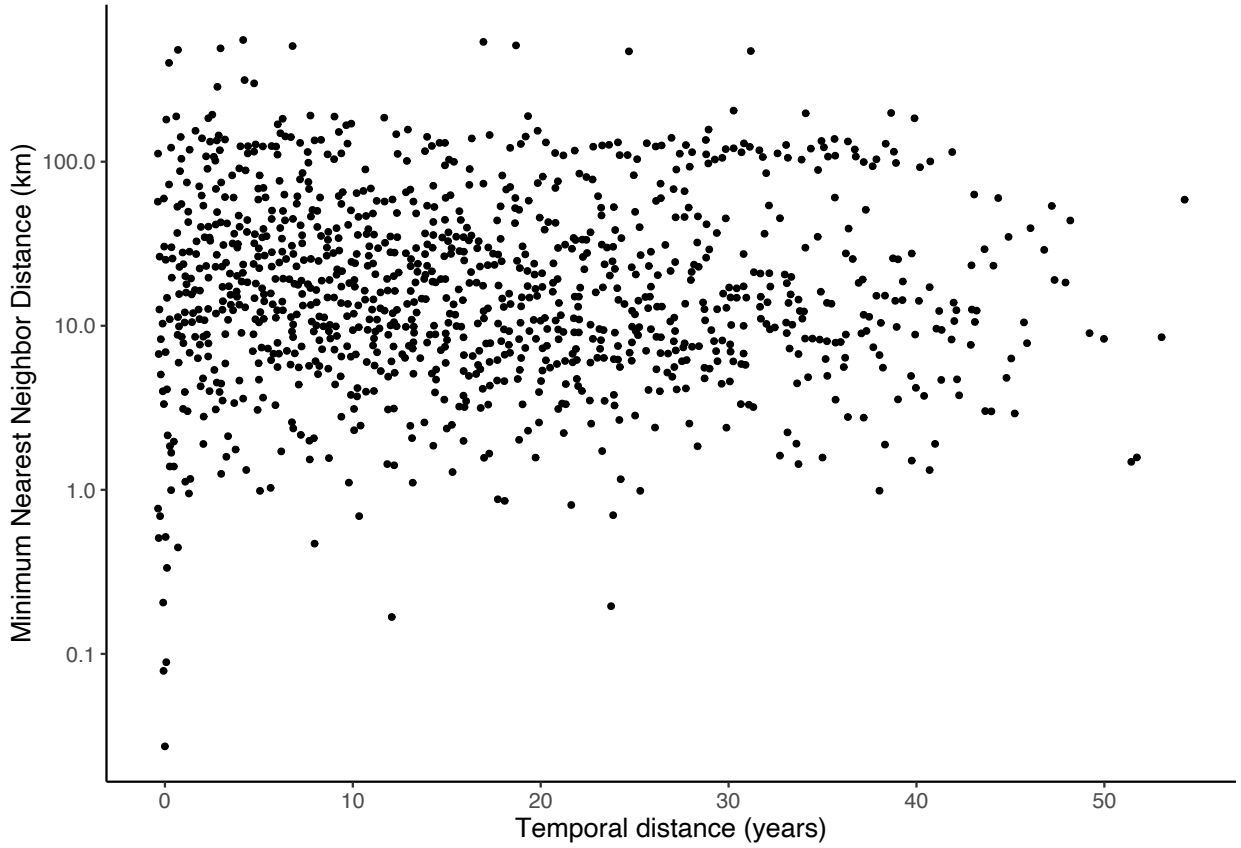
Oriental fruit fly detections in California by year from 1960 through 2014. Light grey bars represent detections in the complete dataset. Dark grey bars represent detections in the outbreak-collapsed version of the dataset.

Figure 1.3.



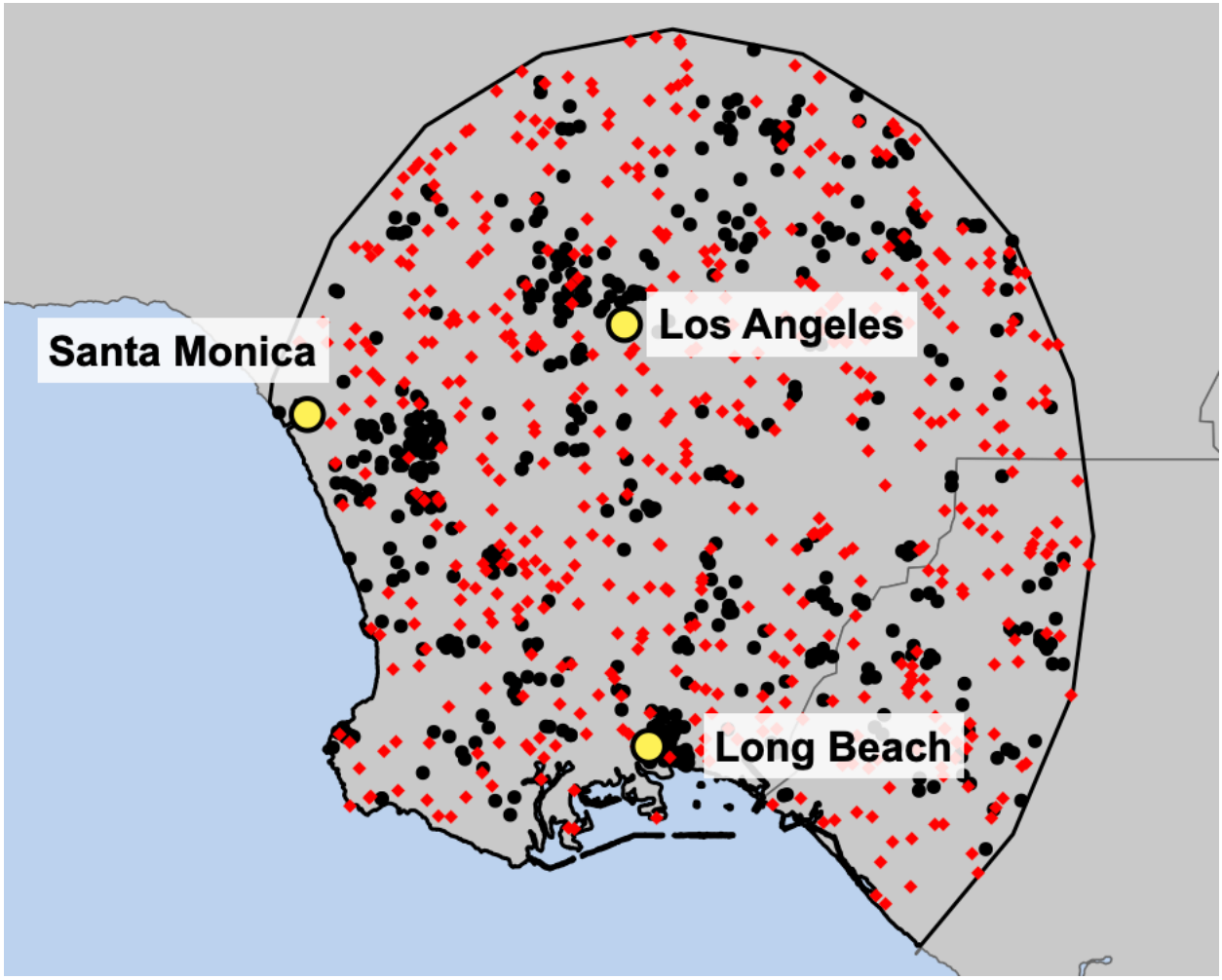
(a) Minimum and (b) median nearest neighbor distances of *Bactrocera dorsalis* fruit flies in California by year. Line colors correspond with the nearest neighbor analysis comparison category. Red lines indicate distances from same-year comparisons; orange lines represent distances from focal year to 1 year prior to comparisons; green lines represent distances from focal year to 3 years prior to comparisons; and blue lines represent distances from focal year to 5 years prior.

Figure 1.4.



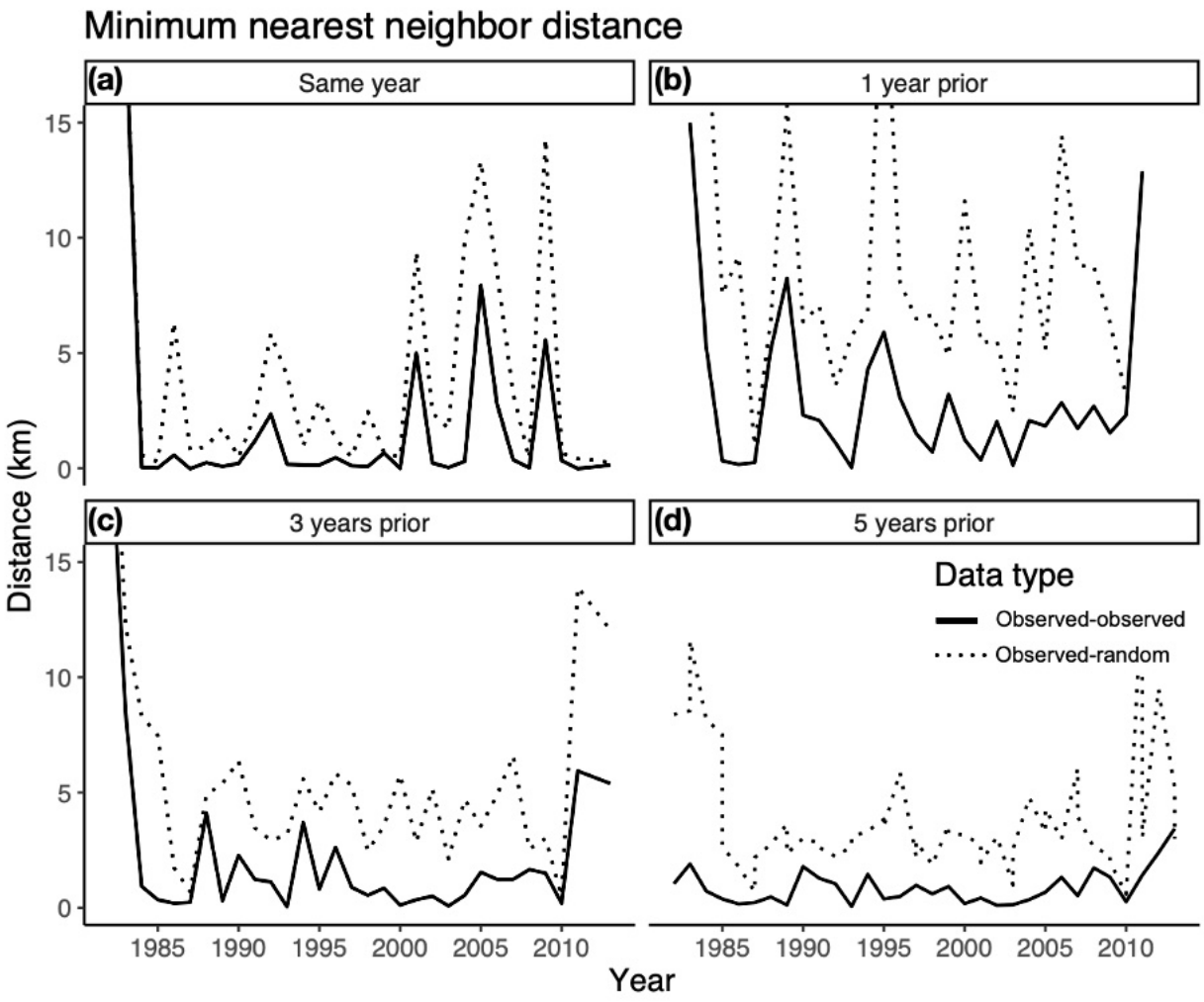
Minimum nearest neighbor distances of B. dorsalis fruit flies in California as a product of temporal distance. Each point represents the smallest nearest neighbor distance found between detections in a focal year and a single other comparison year. The temporal distance in years between the two patterns considered is plotted on the x-axis. The spatial distance in kilometers is plotted on a log10 scale on the y-axis. NNDs are smallest, or most tightly clustered, when detections are temporally close. However, clustering at small distances persists regardless of temporal distance.

Figure 1.5.



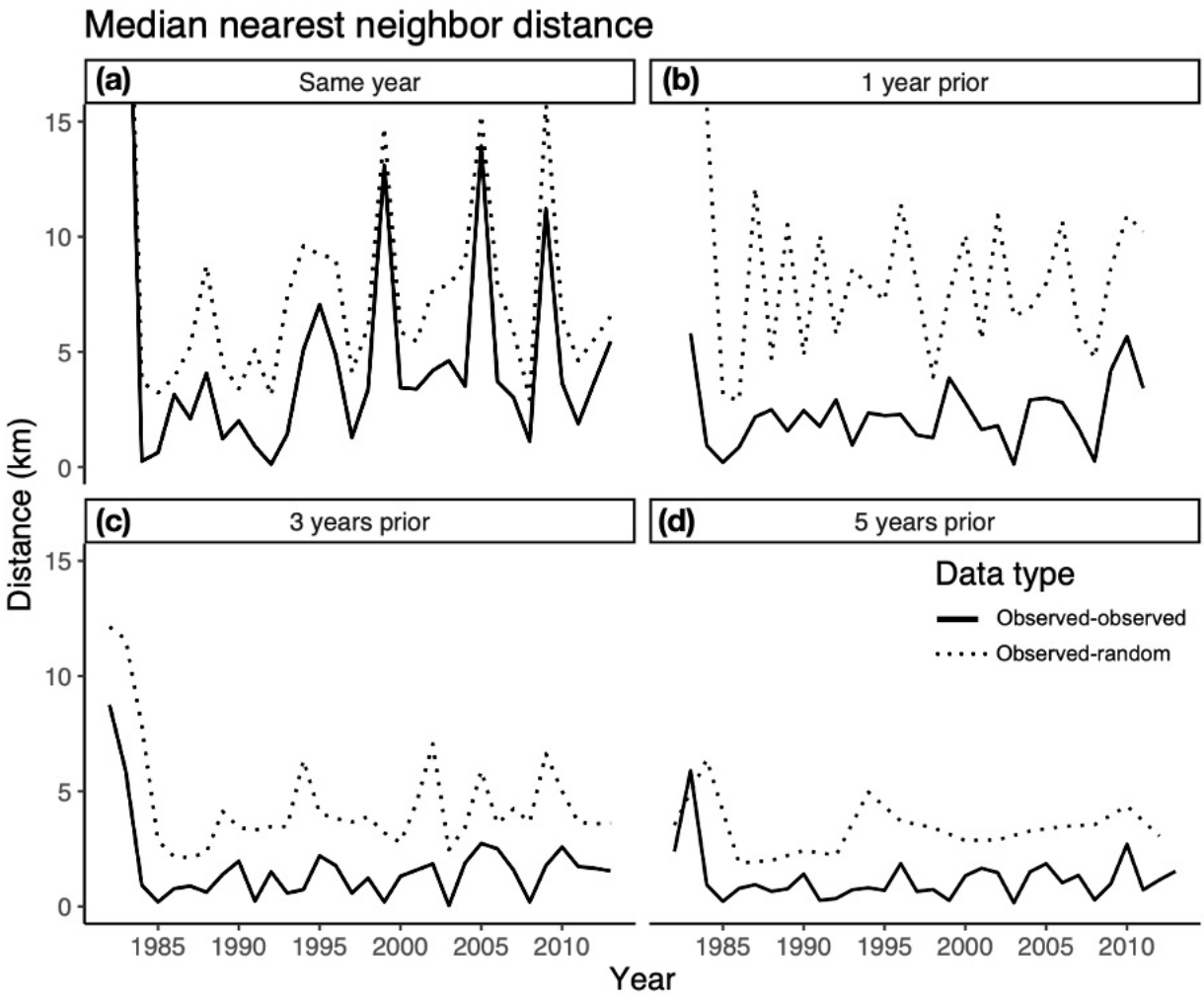
Map depicting the location of the Los Angeles area study region, the observed outbreak-collapsed data (black), the randomly generated data (red), and selected city markers.

Figure 1.6.



Annual minimum nearest neighbor distances of B. dorsalis fruit flies in the southern California study area compared to randomized data. Solid lines represent observed data; dotted lines represent random data. Figure panels reflect the following comparisons: (a) Same-year; (b) focal year to the year prior; (c) focal year to three years prior; (d) and focal year to five years prior. The overall position of the solid line below the dotted line in each comparison indicates that the observed B. dorsalis data are more clustered in space and time (i.e., generates smaller nearest neighbor distances) than would be expected under random distributions.

Figure 1.7.



Annual median nearest neighbor distances of B. dorsalis fruit flies in the southern California study area compared to randomized data. Solid lines represent observed data; dotted lines represent random data. Figure panels reflect the following comparisons: (a) same year; (b) focal year to the year prior; (c) focal year to three years prior; (d) and focal year to five years prior. The overall position of the solid line below the dotted line in each comparison indicates that the observed B. dorsalis data are more clustered in space and time (i.e., generates smaller nearest neighbor distances) than would be expected under random distributions.

Table 1.1.

(A)

Spatial range = 5 km Temporal range = 5 years			Number of future neighbors			Total
			0	1-5	6+	
Number of previous neighbors	0	Count	**334**	**345**	**18**	**697**
		% within previous	48%	49%	3%	100%
		% within future	64%	51%	15%	53%
		% of total	25%	26%	1%	53%
	1-5	Count	**161**	**262**	**75**	**498**
		% within previous	32%	53%	15%	100%
		% within future	31%	39%	64%	38%
		% of total	12%	20%	6%	38%
	6+	Count	**30**	**70**	**24**	**124**
		% within previous	24%	56%	19%	100%
		% within future	6%	10%	21%	9%
		% of total	2%	5%	2%	9%
Total		Count	**525**	**677**	**117**	**1319**
		% within previous	40%	51%	9%	100%
		% within future	100%	100%	100%	100%
		% of total	40%	51%	9%	100%

$X^2 = 95.308$; df = 4; $p < 2.2e\text{-}16$

(B)

Spatial range = 15 km Temporal range = 5 years			Number of future neighbors			Total
			0	1-5	6+	
Number of previous neighbors	0	Count	**89**	**73**	**48**	**210**
		% within previous	42%	35%	23%	100%
		% within future	54%	18%	6%	16%
		% of total	7%	6%	4%	16%
	1-5	Count	**46**	**212**	**299**	**557**
		% within previous	8%	38%	54%	100%
		% within future	28%	53%	40%	42%
		% of total	3%	16%	23%	42%
	6+	Count	**31**	**112**	**409**	**552**
		% within previous	6%	20%	74%	100%
		% within future	19%	28%	54%	42%
		% of total	2%	8%	31%	42%
Total		Count	**166**	**397**	**756**	**1319**
		% within previous	13%	30%	57%	100%
		% within future	100%	100%	100%	100%
		% of total	13%	30%	57%	100%

$X^2 = 280.52$; df = 4; $p < 2.2e\text{-}16$

Chi-squared analysis comparing the number of previous detections within (A) 5km or (B) 15km of a focal detection to the number of future detections within (A) 5km or (B) 15km of the focal detection. Focal points with any previous neighbors are more likely to have neighbors in the future.

Spatiotemporal aggregation analysis reveals which non-native fruit fly populations are likely established in California

Caroline C. Larsen-Bircher, Robert J. Hijmans, James R. Carey

Abstract

Tephritid fruit fly species are some of the most damaging insect agricultural pests. In California, where non-native tephritid species are heavily monitored, annual detections range from frequently-found to rarely-found depending on species. We hypothesize that despite having limited detection data, occasionally-detected species can be still be monitored for spatiotemporal aggregation, an early sign of establishment and will exhibit clustering characteristics intermediate to both rarely- and frequently- detected species. We use two separate spatial aggregation statistics, the *L* function and the O-ring statistic, to determine risk of establishment of six non-native tephritid species: frequently-detected *Bactrocera dorsalis* and *Ceratitis capitata,* and occasionally-detected *Anastrepha ludens*, *Bactrocera correcta*, *Bactrocera zonata*, and *Bactrocera cucurbitae*. In line with previous studies, we find *B. dorsalis* and *C. capitata* to show strong signs of spatiotemporal aggregation. *A. ludens* and *B. correcta* show signs of spatiotemporal aggregation indicating higher establishment risk overall. Spatiotemporal patterns of *B. zonata*, and *B. cucurbitae* generally did not differ strongly from random distributions, thereby posing a lower invasion risk.

Introduction

Non-native tephritid fruit fly species have long been some of the most threatening insect pests to global agriculture. This is especially true in California. California produces more than one third of the country's vegetables and two thirds of the country's fruits and nuts,[1] many of which are potential hosts to tephritid fruit flies.[2]

Since 1900, 17 species of non-native tephritid fruit flies have been detected in California. Of these, most have been found infrequently and in small numbers. 52% of species have had two or fewer individuals detected in California between 2000 and 2014. This lower detection frequency is likely due to low habitat suitability (tropical species in a Mediterranean climate),[3,4] successful post detection control measures by the California Department of Food & Agriculture (CDFA),[5,6] or decreased propagule pressure.[7,8] Such infrequently detected species cannot be analyzed using spatial statistics to estimate establishment. Many of these species have not been detected in the last decade and therefore pose a low risk of establishment. Some non-native tephritids, however, have been detected with great frequency and in great abundance. *Bactrocera dorsalis* (oriental fruit fly), for example, has been detected annually since 1969 and is likely established in population pockets around the state.[9–11] *Ceratitis capitata* (Mediterranean fruit fly), while perhaps the most notorious non-native tephritid species, has been detected in great numbers but with more erratic frequency.[12–14] These frequently detected species have enough spatial data to be analyzed rigorously and have been the subject of many studies.

There are multiple species between these two extremes of frequently and rarely found: species that are detected occasionally, with too many occurrences to be considered eradicated but too few data points for more rigorous statistical analysis necessary to demonstrate establishment. Occasionally-detected species may pose an equally high risk if they have been detected recently or have exhibited a period of repeated detections in a given area, potentially indicating a sub-detectable population. Most of these species have not yet been studied in an ecological context. It is critical to determine which statistical tools will improve our understanding of these species to further understand their establishment risk.

There are many ecological ways to define the risk of any non-native or potentially invasive species[3,1516,17] and of tephritid species in particular.[8,12,18] In this study, we refer to "risk" as the likelihood of a give species to be or become established in a particular area. We know populations can remain sub-detectable for years, so it is critical to take a deep look into the spatial distributions of all detections over time. When spatial analyses include (a) explicit temporal information, (b) a defined study area, and (c) a consideration of recent detection, they can provide a more complete picture of risk than recent detection history alone.

We hypothesize that despite having limited detection data, occasionally-detected species can still be monitored for spatiotemporal aggregation, an early sign of establishment, over time. We posit that these occasionally-detected species will exhibit clustering characteristics intermediate to both rarely- and frequently- detected species. We use spatial statistics to differentiate which, if any, species merit being considered higher risk of establishment based on clustering metrics.

Spatial ecology suggests that spatial statistics are best used in concert due to their varying strengths, sensitivities, and limitations.[19,20,21] We use two separate spatial statistics, the *L* function and the O-ring statistic, to analyze the spatiotemporal patterns of six non-native tephritid species: frequently-detected *Bactrocera dorsalis* and *Ceratitis capitata,* and occasionally-detected *Anastrepha ludens*, *Bactrocera correcta*, *Bactrocera zonata*, and *Bactrocera cucurbitae*. Each of these spatial statistics analyzes degree of clustering among points in a dataset: higher values indicate clustering while lower values indicate no difference from a random distribution. By considering the dataset in 3-year moving window, we can determine aggregation levels in both space and time simultaneously – a greater indication of potential establishment.

Methods

Dataset and Data Selection

The tephritid detection dataset utilized for this study is comprised of 3489 individual detections, their coordinates (longitude, latitude), collection dates (day, month, year), species, sex, life stage, and number of individuals. Over the last century, 17 non-native tephritid species have been detected in California. Detections are made by the CDFA and USDA using approximately five baited traps per square mile across most urban and suburban regions of the state. James Carey and colleagues digitized and curated the detection dataset through 2014.[5,6,8]

We ranked the 17 tephritid species by total detections and selected species with enough cumulative detections, using a minimum threshold of 20 cumulative detections in the greater Los Angeles area. We then divided the species into three tiers: frequently detected, occasionally

detected, and rarely detected. We generated a study area from the cumulative detections of all study species using common methods in geospatial analysis.[4,21,22]

Analysis

All analyses were conducted in R version 3.6.2.[23] Spatiotemporal statistical analyses were primarily performed with the *spatstat*,[24] *onpoint*[25] and *raster*[26] packages. Results were visualized using *ggplot2* package.[27] The dataset is maintained in Microsoft Excel.[28]

Ripley's K(d) and the L function: Like much of point pattern analysis, Ripley's *K* is a distance-based measure: it is calculated based on the distances between events, or in this case points of detection, in a data set.[20,21,29] The *K* function models the distribution of distances between all events, in this case detections, in a given dataset. Unlike other common spatial statistics such as the *F* and *G* functions, which are nearest-neighbor-based, the *K* function provides a glimpse at the overall spatial structure of a given pattern. This statistic considers one event (detection) at a time by taking a ring around that event (detection) at a given radius (*r*) and counting the number of other events (detections) falling inside of that circle. The process is repeated with a slightly larger radius and again until the entirety of the study space has been considered. The next event (detection) is then focused upon and the process repeats. The *K* function is a cumulative function: as the radius increases, it includes all events (detections) inside of the radius. The *L* function includes the same base calculation as Ripley's *K*(*r*) with a linear transformation to make graphical interpretation slightly easier. We used Ripley's isotropic edge correction for all models.[30,31]

The O-Ring Statistic: The O-ring statistic, also referred to as the pair correlation function or neighborhood density function, is another derivative of Ripley's *K*. Similarly, to the *K* and *L* functions, it provides a statistical estimation of the spatial structure of a given point pattern. Unlike the *K* and *L* functions, the O-ring statistic is a probability density function: it separates the distances between event pairs into spatial bins, rather than considering them cumulatively.[20,21,29]

Monte Carlo Testing: Tests of significance for spatial point patterns most commonly use Monte Carlo procedures. Monte Carlo simulation of a spatial pattern involves using the same statistical parameters with a simulated data set representing the null hypothesis, which here is a random distribution.[21] Monte Carlo procedures can be used to test hypotheses other than complete spatial randomness (CSR); however, we have chosen a null of random distributions for this analysis. Future studies will explore the variable impact of more complex temporally explicit null distributions. Statistical rejection limits are based on simulation envelopes, generated by each of the 99 iterations of the null test. Each of the 99 simulations includes a distinct randomly generated spatial pattern within the same study area polygon as the observed data sets and containing the same number of events (detections) as each comparable cumulative or annual time slice as the observed data set. In both versions of both statistical tests, if the observed data is higher than the simulation envelope, the pattern is clustered, or aggregated.[20,21] If the observed data falls inside the simulation envelope, the pattern is random. If the observed data falls below the simulation envelope, it is regular, or over-dispersed.

Time: For each statistical test, we incorporated time using two approaches: cumulative years and 3-year windows. We first considered time cumulatively, with all years of detections combined into a single pattern. In the 3-year window analyses, data was considered in moving windows of time. For example, the combined detection patterns for 1990-1992 would be analyzed as a unit, followed by the combined patterns for 1991-1993. Using a moving window framework can help illuminate patterns in sparser data sets by smoothing the data over time. Detections in close spatial and temporal proximity are highlighted in the analysis, which follows ecologically as detections close together in both space and time are more likely to be the result of a small population. Each window results in a similar output to the cumulative time tests. To display the statistical information in a consolidated and more easily digestible form, we have used quantum plots.[25,32]

Results

We ranked the 17 tephritid species and divided them into categories (Table 2.1). *B. dorsalis* (1421 total detections; 953 study area detections) and *C. capitata* (1396 total; 783 study area) comprise the frequently detected species. *A. ludens* (437 total; 124 study area), *B. correcta* (139 total; 89 study area), *B. zonata* (68 total; 37 study area), *and B. cucurbitae* (28 total; 23 study area) make up the occasionally detected tier. The remaining 14 species did not pass the 20-detection minimum threshold and fell into the rarely detected category. These species did not have enough spatial data to be considered in the analyses.

The selected species broadly overlap in range, most noticeably in the greater Los Angeles Area. We therefore confined our analyses to data in this study region, utilizing detections inside a

single polygon boundary nested within the detection ranges of all six species. The study area polygon was generated using a combination of buffer functions on the observed detected points (Figure 2.2).

Cumulative years analysis

L function: When considered cumulatively, each of the six tephritid species shows significantly higher levels of aggregation than the simulation envelope at 100% of distances *r* (Figure 2.3). This difference is most pronounced with the two frequently-detected species, *B. dorsalis* and *C. capitata*. *A. ludens*, *B. correcta*, and *B. zonata* also show statistical differentiation from the simulation envelope, though the separation is narrower. *B. cucurbitae,* the least abundant of the six species, hovers most closely to the simulation envelope at higher distances *r*, though still shows higher aggregation than random at small distances. The simulation envelopes of the less-frequently detected species are much more variable due to the limited number of detections in each dataset, and thereby in each Monte Carlo null simulation.

O-ring statistic: Similar to the cumulative year tests for the *L* function, observed spatial patterns from frequently detected species using the O-ring statistic show substantial difference from the Monte Carlo simulation envelope, especially at small distances *r* (Figure 2.4). *B. dorsalis*, *C. capitata*, and *A. ludens* show the greatest levels of aggregation and difference from the simulation envelope. The observed *O(r)* values for *B. zonata* and *B. cucurbitae*, however, fall generally within the simulation, indicating no difference from a random distribution at most distances *r*.

3-year annual window analysis

L function: The results of the 3-year window *L* function tests were far more variable. Frequently-detected species show solid spatiotemporal aggregation at all distances *r* (Figure 2.5). *B. dorsalis* shows the strongest aggregation over time, in line with previous studies. Other species show more temporal variation. For example, *B. correcta* shows little aggregation at small distances *r* across time but has extremely variable aggregation at larger scales. In recent years, *B. correcta* shows only some aggregation and few neighbors. The lack of detection data for *B. zonata* and *B. cucurbitae* is evident in the sparsely filled bars.

O-ring statistic: The annual analysis of the O-ring statistic again shows the most frequently-detected species presenting strong signals of aggregation at small distances *r* across time (Figure 2.6). *B. correcta* shows some spatiotemporal aggregation, but most recently patterns were no different from random. *B. dorsalis, C. capitata, A. ludens,* and *B. zonata* all show evidence of overdistribution, or regularity at medium and large distances.

Discussion

Ecological implications

Spatial statistics are critical tools for revealing underlying dynamics of ecological populations. The results of our study support our hypotheses that even with limited amounts of data, occasionally detected tephritid species can be shown to have spatiotemporal clustering. Analyses of frequently-detected species confirm the result of other studies: these species are a continued high risk and most likely established in the study area. Occasionally-detected species vary in their risk assessment: of the four, B. correcta poses the highest risk of establishment. Further, when used correctly, these results can be used to track species risk in a particular area.

Frequently-detected species: Overall, *B. dorsalis* is the most abundantly detected non-native tephritid species in California. The species is found annually and has been demonstrated to be likely established in prior studies. On this information alone, *B. dorsalis* is a high-risk species. This is confirmed through all four statistical tests. The cumulative years *L function* and *O-ring* statistic both indicate that *B. dorsalis* detections aggregate in space: new detections are likely to be near-by to older detections and aggregated in patterns that are statistically distinct from random distributions within the study area (Figure 2.3; Figure 2.4). The 3-year window *L* function shows that *B. dorsalis* detections are aggregated up through most all distances *r* (0-0.20) across the study area (Figure 2.5). The 3-year O-ring analysis provides a bit more structural detail with the same overall result: solid clustering signals at small distances of *r* (Figure 6). Considered together, these results align with those of previous studies suggesting population persistence, and likely establishment, of *B. dorsalis* in the Los Angeles area.

Perhaps the most recognized of the California non-native tephritid species, *C. capitata* has been a primary focus of the CDFA for decades. It is comparable to *B. dorsalis* in overall detections, but the annual frequency is more variable, showing up 75% of the last 15 years as opposed to 100% for *B. dorsalis* (Table 2.1; Figure 2.1). This is reflected in multiple spatial analyses. For example, the cumulative years O-ring analysis, the observed output for *C. capitata* is far more variable in terms of amplitude than *B. dorsalis* (Figure 2.4). Whereas *B. dorsalis* patterns show overall clustering at mostly small distances *r*, the observed curves for *C. capitata* approach the MC simulation envelope more erratically. This may be due to the fact that historically there have been multiple extreme outbreak years, which skew the cumulative years analyses. Both 3-year window models show that *C. capitata* populations within the study area variably show

aggregation and random distributions (Figure 2.5; Figure 2.6). The most recent year of data changes a three-year streak of largely random patterns to show aggregation at small distances *r*.

Occasionally-detected species: Though *A. ludens* was detected in California before any of the other five species (1954), its total detections are merely a third of those for *B. dorsalis* and *C. capitata* (Table 2.1). While the prior two frequently-detected species have a signature of consistent aggregation at small distances *r*, the cumulative years analyses for *A. ludens* indicate more aggregation at mid-level distances (Figure 2.3; Figure 2.4). Three-year window analyses show that the most recent *A. ludens* detections have been clustered (Figure 2.5; Figure 2.6). As there have been no *A. ludens* detections in the study area since 2008, the possibility of an established population is progressively less likely.

B. correcta was first found in the state in 1986 and since then 139 individuals have been detected (Table 2.1). This species has been found 14 out of the last 15 years, making it a higher risk species for future detections. In both cumulative years tests, the distributions of *B. correcta* most closely resembles those of *B. dorsalis* but with a lower abundance: high levels of clustering at small distances *r* with a relatively stable taper (Figure 2.3; Figure 2.4). The 3-year window analyses show that unlike most other species, the smallest distances *r* indicate a random distribution, likely driven by the lower overall abundance of detections (Figure 2.5; Figure 2.6). By $r = 0.02$, the pattern show aggregation. Most recent years indicate more random distributions than aggregated. However, due to the consistent detections within the study area and the

aggregation signals, though variable, future detections of *B. correcta* should be carefully considered to determine if these populations may already be established in the area.

The last two species, *B. zonata* and *B. cucurbitae*, have the fewest overall detections and highly variable annual distributions (Table 2.1; Figure 2.1), only being found in 6 (*B. zonata*) or 4 (*B. cucurbitae*) of the last 15 years. The cumulative years analyses for both species show little overall differential from the null simulation windows (Figure 2.3; Figure 2.4). The 3-year window analyses similarly show that distributions are variable: generally random at middle and large distances *r*, but variably aggregated at small distances *r* (Figure 2.5; Figure 2.6). *B. zonata* had a large temporal gap with no detections within the study area. Individuals were most recently detected in 2013 (showing up as bars for 2013 and 2014 in the 3-year window analyses) but did show aggregation at small distances. With such few total detections, it is difficult to determine whether *B. zonata* is a high risk (established) species. The model results, limited detections, and annual variation do not indicate current establishment, though it is a possibility. *B. cucurbitae*, on the other hand, has not been detected in the study area since 2010. Both 3-year window *L* function and O-ring statistic for 2010 show no difference from the random distribution simulation envelope, and preceding years were similar. This suggests that *B. cucurbitae* is less likely to be established in the study area. Of course, any introduction of non-native tephritids can potentially lead to (or signal) population establishment or outbreak, so some amount of risk is associated with even a single fly detection.

Management implications

Spatial statistics can be a useful tool in management if used correctly. Spatiotemporal analyses have strengths and weaknesses, and thus should be used collectively rather than alone. Both cumulative years models give important information about the structure of patterns overall, though are likely less useful for the rarer species. The *L* function seems to be less useful in a test against the null hypothesis as it may overestimate aggregated versus the random simulations. However, the initial shapes of the curve are important. Any rapid increase along the x-axis indicates scales of clustering. The O-ring statistic also produces structural information, seems less influenced by abundance, and perhaps is simpler to interpret. The downside of these two analyses is of course that they are temporally cumulative and have no information on change over time. This doesn't mean it void of ecological significance: we've seen, especially in the early years of *B. dorsalis* and *C. capitata*, that detections close together in space but years apart in time can be a part of a sub-detectable population.

We found the 3-year window to be the most useful tool for determining recent risk, for frequently and occasionally occurring species, especially when paired with individual O-ring diagrams. The O-ring statistic in particular is easy to interpret visually in terms of identifying spatiotemporal scale of aggregation. Though these tests do show temporal variation, they do not indicate how much each pattern differs from the simulation envelope, a benefit of the cumulative year tests.

Acknowledgements

We would like to thank Dr. Jay Rosenheim, Dr. Ash Zemenick, and Dr. Jenny VanWyk for helpful feedback on the manuscript. We thank the Carey and Hijmans lab for advice. This material is based on work supported by National Science Foundation Graduate Research Fellowship Grant. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation. Specific author contributions: CCL conceived the study, analyzed data, and wrote the initial draft. RJH assisted with coding and analysis. JRC provided the dataset, expertise on the study species, and contributed substantial feedback at each stage of the project.

Works Cited

1. California Department of Food & Agriculture. *California Agricultural Statistics Review*. (2020).
2. Bateman, M. The ecology of fruit flies. *Annu. Rev. Entomol.* **17**, (1972).
3. Davis, M. A. *Invasion biology*. (Oxford University Press, 2009).
4. Gaston, K. J. *The Structure and Dynamics of Geographic Ranges*. *Oxford Series in Ecology and Evolution* (2003). doi:10.2167/jost191b.0
5. Gilbert, A. J., Bingham, R. R., Nicolas, M. A. & Clark, R. A. *Insect Trapping Guide*. (2013).
6. California Department of Food & Agriculture & U.S. Department of Agriculture. Exotic Fruit Fly Regulatory Resonse Manual. (2001). doi:10.1093/infdis/jit776
7. Simberloff, D. The Role of Propagule Pressure in Biological Invasions. *Annu. Rev. Ecol. Evol. Syst.* **40**, 81–102 (2009).
8. Papadopoulos, N. T., Plant, R. E. & Carey, J. R. From trickle to flood: the large-scale, cryptic invasion of California by tropical fruit flies. *Proc. Biol. Sci.* **280**, 20131466 (2013).
9. Stephens, A. E. A., Kriticos, D. J. & Leriche, A. The current and future potential geographical distribution of the oriental fruit fly, Bactrocera dorsalis (Diptera: Tephritidae). *Bull. Entomol. Res.* **97**, 369–378 (2007).

10. Froerer, K. M. *et al.* Long-distance movement of Bactrocera dorsalis (Diptera: Tephritidae) in Puna, Hawaii: How far can they go? *Am. Entomol.* **56**, 88–95 (2010).

11. Carey, J. R., Papadopoulos, N. T. & Plant, R. Tephritid pest populations oriental fruit fly Outbreaks in California: 48 consecutive years, 235 Cities, 1,500 detections-and counting. *Am. Entomol.* **63**, 232–236 (2017).

12. Carey, J. R. Establishment of the Mediterranean Fruit Fly in California. *Science (80-.).* **253**, 1369–1373 (1991).

13. Papadopoulos, N. T., Katsoyannos, B. I., Carey, J. R. & Kouloussis, N. A. Seasonal and Annual Occurrence of the Mediterranean Fruit Fly (Diptera: Tephritidae) in Northern Greece. *Ann. Entomol. Soc. Am.* **94**, 41–50 (2001).

14. Diamantidis, A., Carey, J. R. & Papadopoulos, N. T. Life-history evolution of an invasive tephritid. *J. Appl. Entomol.* **132**, 695–705 (2008).

15. *Encyclopedia of Biological Invasions*. (University of California Press, 2011).

16. Robinet, C. & Roques, A. Direct impacts of recent climate warming on insect populations. *Integr. Zool.* **5**, 132–42 (2010).

17. Biber-Freudenberger, L., Ziemacki, J., Tonnang, H. E. Z. & Borgemeister, C. Future risks of pest species under changing climatic conditions. *PLoS One* **11**, 1–17 (2016).

18. Zhao, Z. *et al.* Life table invasion models: spatial progression and species-specific partitioning. *Ecology* **100**, 1–11 (2019).

19. Diggle, P. J. On Parameter Estimation and Goodness-of-Fit Testing for Spatial Point Patterns. *Biometrics* **35**, 87–101 (1979).

20. Perry, G. L. W., Miller, B. P. & Enright, N. J. A comparison of methods for the statistical analysis of spatial point patterns in plant ecology. *Plant Ecol.* **187**, 59–82 (2006).

21. O'Sullivan, D. & Unwin, D. J. *Geographic Information Analysis: Second Edition*. (John Wiley & Sons, Inc, 2010). doi:10.1002/9780470549094

22. Diggle, P. J. *Statistical analysis of spatial and spatio-temporal point patterns, third edition*. (Taylor & Francis Group, LLC, 2013). doi:10.1201/b15326

23. R Core Team. R: A Language and Environment for Statistical Computing. (2019).

24. Baddeley, A., Rubak, E., Turner, R. & Raton, B. Spatial Point Patterns: Methodology and Applications with R. *J. Stat. Softw.* **75**, (2016).

25. Hesselbarth, M. H. K. onpoint: Helper functions for point pattern analysis. (2020).

26. Hijmans, R. J. raster: Geographic Data Analysis and Modeling. (2019).

27. Wickham, H. *ggplot2: Elegant Graphics for Data Analysis*. (Springer-Verlag New York, 2016).

28. Microsoft Corporation. Microsoft Excel. (2019).

29. Velázquez, E., Martínez, I., Getzin, S., Moloney, K. A. & Wiegand, T. An evaluation of the state of spatial point pattern analysis in ecology. *Ecography (Cop.).* **39**, 1042–1055 (2016).

30. Ohser, J. On estimators for the reduced second moment measure of point processes. *Ser. Stat.* **14**, 63–71 (1983).

31. Diggle, P. J. *Statistical analysis of spatial and spatio-temporal point patterns, third edition*. (Taylor & Francis Group, LLC., 2013). doi:10.1201/b15326

32. Esser, D. S., Leveau, J. H. J., Meyer, K. M. & Wiegand, K. Spatial scales of interactions among bacteria and between bacteria and the leaf surface. *FEMS Microbiol. Ecol.* **91**, 1–13 (2015).

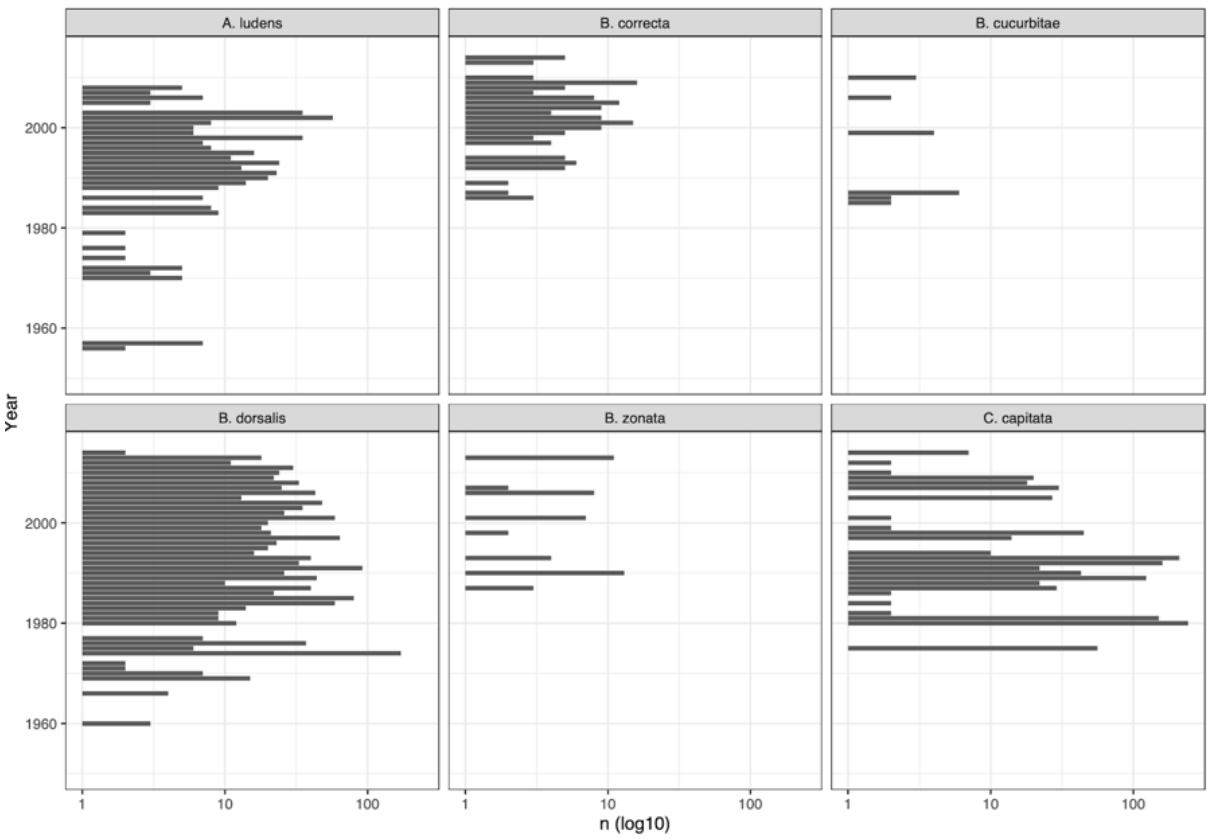
Figures

Table 2.1.

Detection Tier	Species	Total Detections	First Year Detected	Most Recent Year Detected	Total Detections since 2000	Number of Years Detected 2000-2014 (Frequency)
Frequent	***B. dorsalis***	***1421***	***1960***	***2014***	***474***	***15 (100%)***
	C. capitata	***1396***	***1975***	***2014***	***121***	***11 (73%)***
Occasional	***A. ludens***	***437***	***1954***	***2008***	***168***	***8 (53%)***
	B. correcta	***139***	***1986***	***2014***	***102***	***14 (94%)***
	B. zonata	***68***	***1984***	***2013***	***31***	***6 (40%)***
	B. cucurbitae	***28***	***1956***	***2010***	***8***	***4 (27%)***
Rarely	B. scutellata	16	1987	2010	12	2 (13%)
	A. striata	11	1909	1998	0	0
	B. albistrigata	10	2008	2009	10	2 (13%)
	A. obligua	8	1967	2005	2	2 (13%)
	A. suspensa	7	1983	2007	1	1 (7%)
	B. tryoni	2	1985	1991	0	0
	D. bivittatus	1	1987	1987	0	0
	A. serpentina	1	1989	1989	0	0
	B. facialis	1	1998	1998	0	0
	B. latifrons	1	1998	1998	0	0
	A. obliqua	1	2000	2000	1	1 (7%)

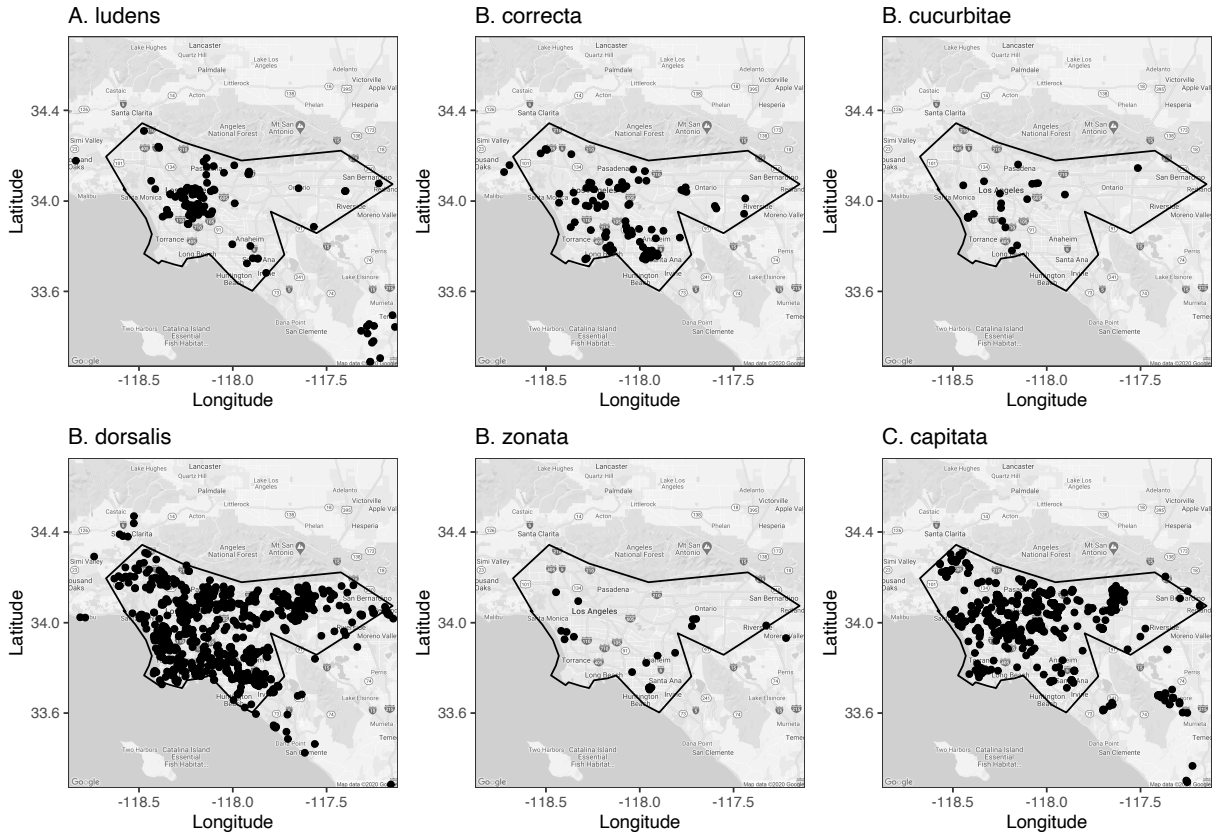
Detection history of 17 non-native tephritid species in California from 1900 through 2014. Species in bold are the six focal species of this study. Species in light grey are those that have not been detected in the last 15 years (2000-2014).

Figure 2.1.



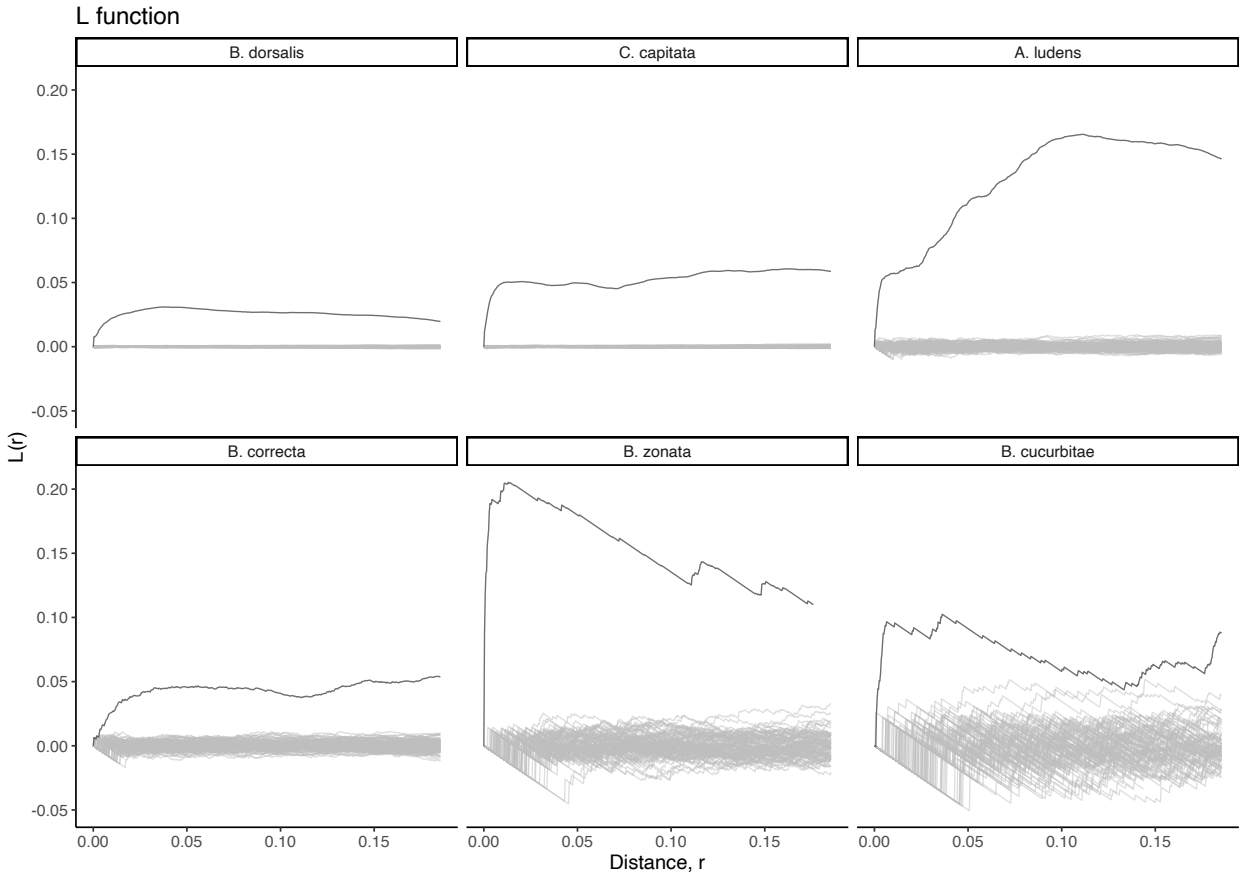
Annual detections of the top 6 most frequently detected non-native tephritid species in California. The x-axis is log10 scale of *n*, the total annual abundance, to account for the highly variable detection amounts.

Figure 2.2.



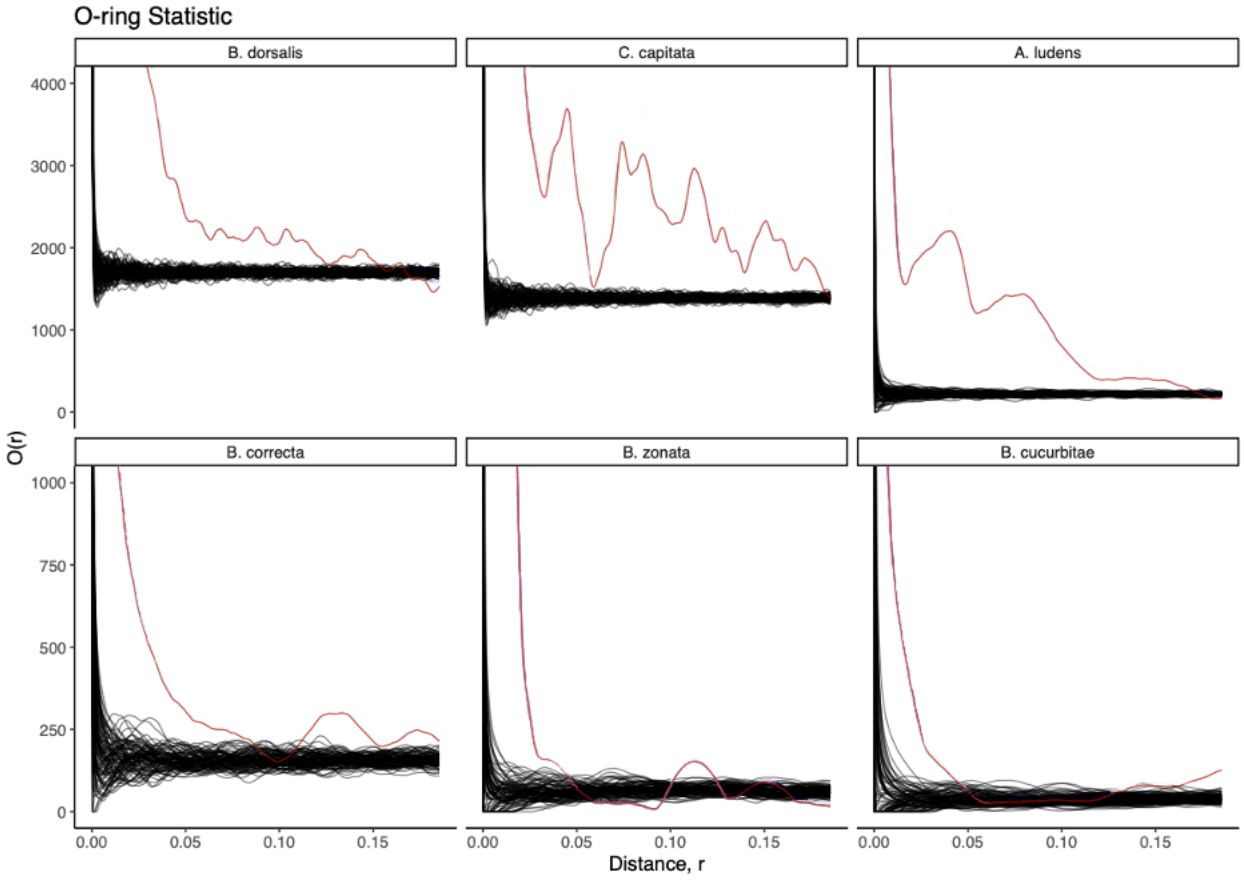
Cumulative detections of the top 6 most frequently detected non-native tephritid species in the Los Angeles area of California. Individual detections are represented by black circles. The study area polygon is shown by the black line in each window.

Figure 2.3.



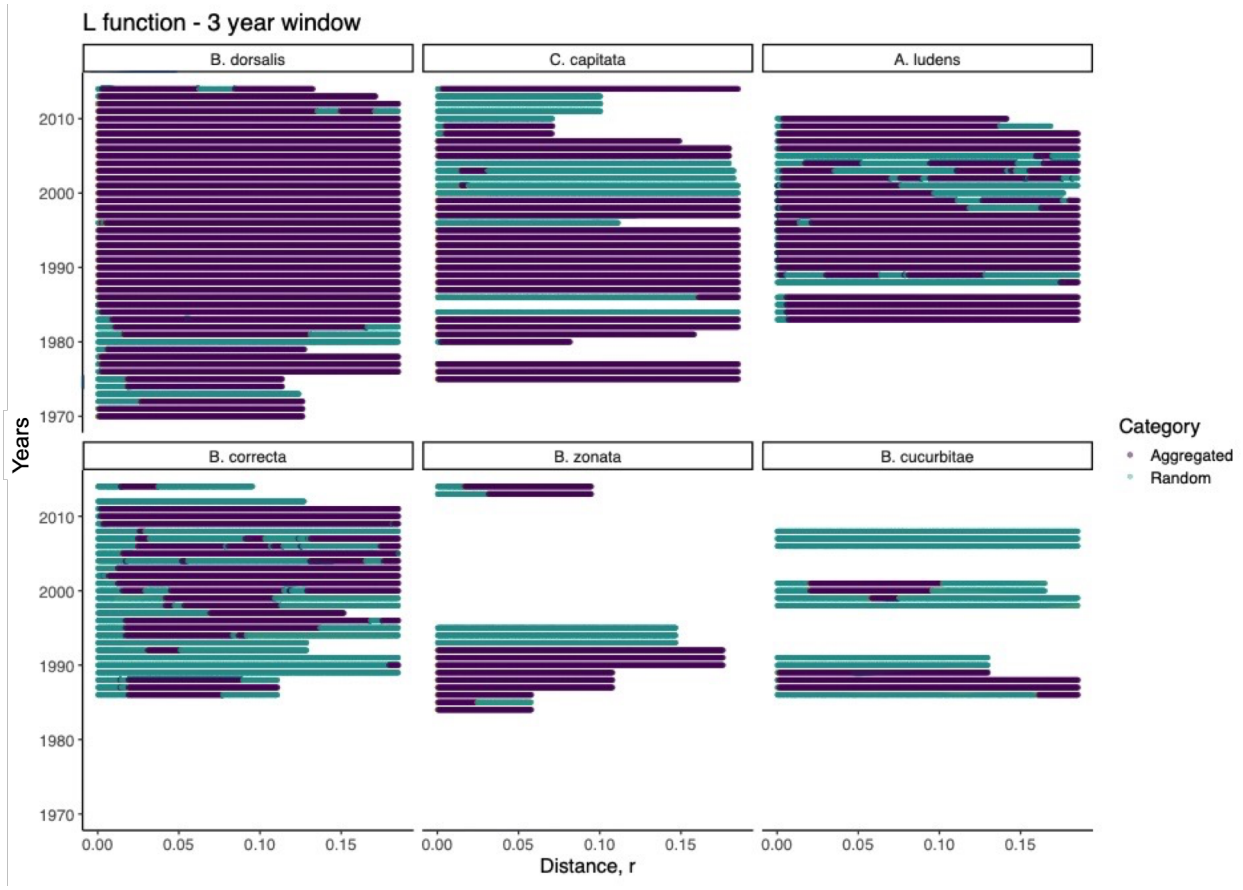
L functions for the top 6 most frequently detected non-native tephritid species in the Los Angeles area of California. Distance *r* is the radius of the circle surrounding an event (detection). *L(r)* described the average number of events inside a circle of radius *r*, with a linear transformation. The black line in each panel represents *L(r)*for the observed cumulative detection data for each species. Higher values of L(*r*) denote higher levels of spatial aggregation. The grey lines in each panel represent *L(r)*for the simulated Monte Carlo simulations.

Figure 2.4.



O-ring statistics for the top 6 most frequently detected non-native tephritid species in the Los Angeles area of California. Distance *r* is the radius of the annuli (ring) surrounding an event (detection). *O(r)* described the average number of events inside an annulus of radius *r*. The red line in each panel represents *O(r)* for the observed cumulative detection data for each species. Higher values of *O(r*) denote higher levels of spatial aggregation. The black lines in each panel represent *O(r)*for the simulated Monte Carlo simulations.

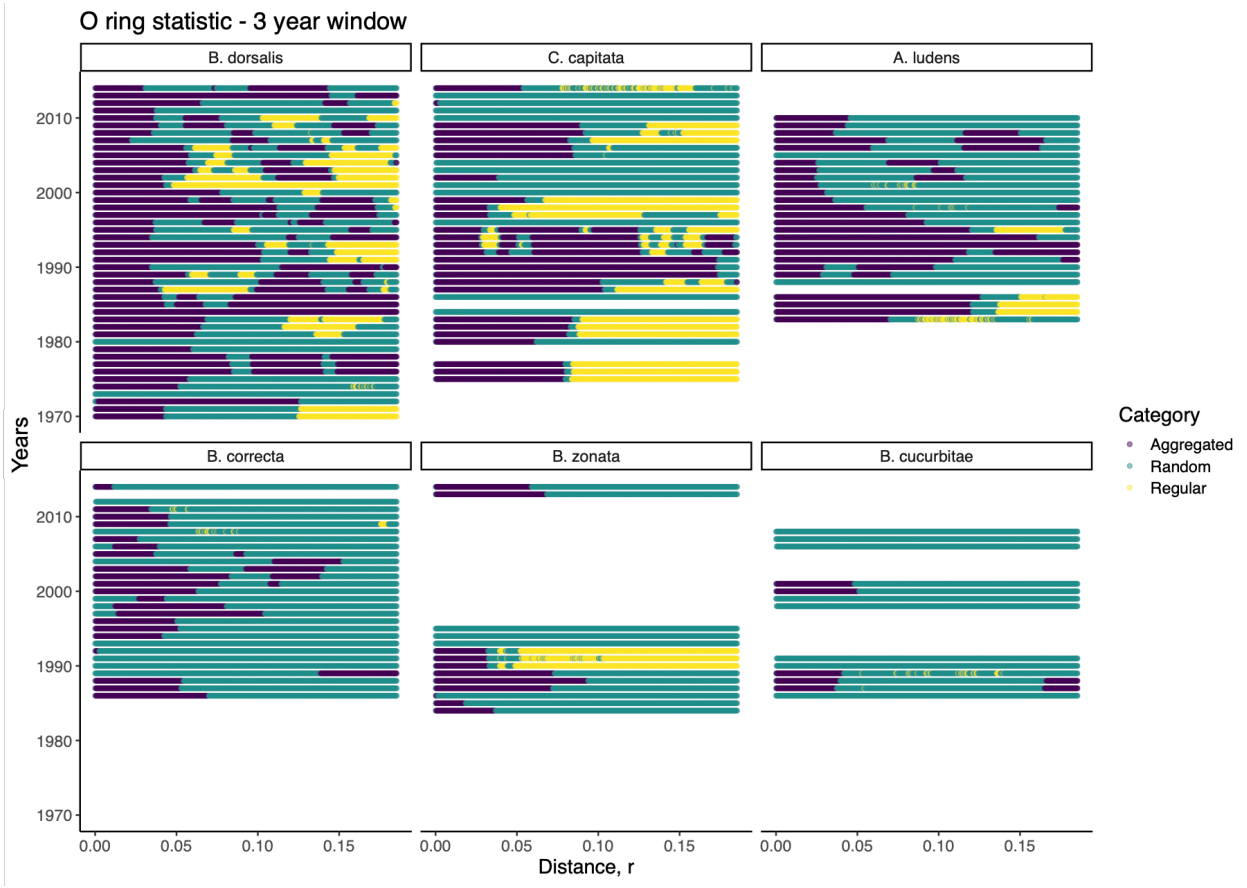
Figure 2.5.



L functions for the top 6 most frequently detected non-native tephritid species in the Los Angeles area of California, separated by year. Distance *r* is the radius of the circle surrounding an event (detection). *L(r)* described the average number of events inside a circle of radius *r*, with a linear transformation. Each row represents detections from a three-year window of time, listed as the first of the three years (e.g., 1990 represents 1990-1992). Purple portions of each bar represent an aggregated distribution, i.e., the observed L*(r)* value is greater than the Monte Carlo simulation envelope. Green portions of each bar represent a random distribution, i.e., the observed L*(r)* value falls inside the Monte Carlo simulation envelope. If no bar is present for a given year or distance *r*, there were no neighbors in the temporal window in that spatial bin. Short bars that are present

at small distances but disappear at larger distances indicate that near neighbors were present, but no neighbors were found at higher distances.

Figure 2.6.



O functions for the top 6 most frequently detected non-native tephritid species in the Los Angeles area of California, separated by year. Distance *r* is the radius of the annuli (ring) surrounding an event (detection). *O(r)* described the average number of events inside an annulus of radius *r*. Each row represents detections from a three-year window of time, listed as the first of the three years (e.g., 1990 represents 1990-1992). Purple portions of each bar represent an aggregated distribution, i.e., the observed *O(r)* value is greater than the Monte Carlo simulation envelope. Green portions of each bar represent a random distribution, i.e., the observed *O(r)* value falls within the Monte Carlo simulation envelope. Yellow portions of each bar represent a regular distribution, i.e., the observed *O(r)* value falls below the Monte Carlo simulation envelope. If no bar is present for a given year or distance *r*, there were no neighbors in the temporal window in that spatial bin.

Identifying drivers of detections of *Bactrocera dorsalis*, an introduced fruit fly in Los Angeles, California

Caroline C. Larsen-Bircher, Robert J. Hijmans, James R. Carey

Abstract

Non-native invasive species, including insect pests, can cause severe damage to human health, ecosystems, and agriculture. Invasive insect pest populations are influenced by many ecological variables, such as climate and host plant occurrence; however, human-related activities, such as land development, population density, and transportation are known to additionally modify habitat conditions. This suggests that the inclusion of human-related processes may improve our understanding of the spatial distributions of costly and elusive invasive pests. In this study, we examine the distribution of *Bactrocera dorsalis* fruit flies, damaging agricultural pests, in the Los Angeles, California area. Using a combination of ecological, human, and *B. dorsalis* population variables related to mechanisms of tephritid introduction and establishment, we identified the strongest explanatory variables of detections using random forest and logistic regression. Primarily, we find that proximity to prior detections was an important indicator of future detections, supporting the assertion that *B. dorsalis* is established in California. However, some anthropogenic variables related to transportation and demographics, such as the locations of bus stations, were strongly associated with *B. dorsalis* occurrence, indicating that human

modification of landscapes may shape of this invasive species' distribution, beyond environmental variables alone. Together, these results may improve management and detection of *B. dorsalis* in its introduced range and increase our understanding of the deeper dynamics driving these populations.

Introduction

Understanding current distributions and predicting future distributions of non-native or pest species is a critical component of successful management.[1,2] Tephritid fruit flies are one of the most heavily monitored agricultural pest species worldwide. In fruit-eating species, females oviposit under the skin of fruit and the larvae consume the fruit flesh until pupation, thus ruining the plant for human consumption. Though host plants are often citrus, stone fruit, and cucurbits; most fruits and many vegetables have a tephritid species that consider it a host.[3–5] In California, an agricultural economy of global importance, non-native tephritids have been monitored and managed since the early 1900s, given the significant damage they can potentially cause.[6–10] To date, 17 non-native tephritid species have been found in urban and suburban areas of throughout the state. Urban areas are important to this agricultural pest because California agriculture land is kept tephritid-free through intensive preventative management: the populations of concern are persisting in developed areas of the state where human-supported backyard fruit trees and vegetable gardens increase the potential niche of the species.

Bactrocera dorsalis, the oriental fruit fly, is a tropical species native to Southeast Asia. Its range has now extended to at least 65 countries, and as an introduced species, it is a dominant global agricultural threat.[11–14] Its larvae feeds on 400+ documented host plant (plants in the tomato,

squash, and citrus families are particularly vulnerable)[5] and is an important pollinator in its native range. There have been detections of *B. dorsalis* in California every year since 1969. This species is especially abundant in the Los Angeles area despite vigilant monitoring and intervention methods by the California Department of Food and Agriculture.[6,7,10] It appears that *B. dorsalis* has established in the state in small pockets that may shift spatially over time.[15,16] Management of *B. dorsalis* and other non-native tephritid species, found using on-the-ground monitoring, relies on a policy of 'detect and eradicate'. Subsequent detections outside of a small spatiotemporal window are considered new introductions from importation of goods or travel. Populations rarely rapidly increase due to patchy host availability and these post-detection control measures; however, complete eradication, which relies on a 100% success rate, is quite difficult.[15,17–19]

Many models, both inferential and predictive, focus on the impact of bioclimatic variables on tephritid populations, which can be critical for understanding species ranges.[20,21] However, anthropogenic factors can modify existing habitat conditions by maintaining host plants outside of their natural bioclimatic ranges or promoting microclimates due to tree cover or urban heat island effects.[22–27] For instance, populations of tephritids in California may be more likely to persist in backyard fruits trees and home gardens, as compared to the undeveloped landscapes of California that lack suitable hosts. Thus, variables such as degree of developed land cover and human population density may help us understand detection distributions. Evidence supports the hypothesis of *B. dorsalis* establishment in Los Angeles, though monitoring and intervention methods operate on the assumption of frequent repeat introductions.

International importing of goods or civilian or military travel from regions with well-established *B. dorsalis* populations have been cited as possible pathways of *B. dorsalis* introduction. It has been frequently suggested that Los Angeles residents traveling to and from countries with non-native fruit flies are a primary source of new introductions. *B. dorsalis* is a wide ranging tephritid species, endemic to southeast Asia and established elsewhere, most notably in Hawaii. The first line of defense against introduction of new pest species from agricultural imports are heavily regulated inspections. If importation is the primary ongoing pathway for *B. dorsalis* detections via new introductions, detections may correlate with distance to transportation hubs such as airports, ports, and freight transfer facilities or with census tracts with a higher proportion of residents identifying as Asian or Hawaiian compared to other identities. If travel is a primary re-introduction pathway for new detections, public or military airports may significantly predict detections.

- In this study, we use a long-term dataset of *B. dorsalis* detections of to examine the biophysical and anthropogenic drivers of *B. dorsalis* detections, testing variables that may shape both tephritid reintroduction and populations persistence following invasion (Table 3.1). We specifically explore the following hypotheses: (1) *B. dorsalis* occurrence is correlated with proximity to historical detections and not with transportation pathways, indicating that within the urban matrix of the greater Los Angeles area, *B. dorsalis* is established in small, difficult-to-detect populations, rather than through repeated introductions; (2) Anthropogenic factors are significantly correlated with *B. dorsalis* detection and improve model performance, indicating that human modification of landscapes may dictate habitat suitability, beyond biophysical factors alone.

We used two statistical inference methods, the random forest algorithm and logistic regression, to identify important factors of *B. dorsalis* occurrence. Random forest is a regression and classification algorithm frequently used with environmental and biological data that lends well to exploring complex interactions among variables, especially with small sample sizes.[28–30] Logistic regression provides easily-interpretable quantitative estimates of effect size and direction. Together, these statistical approaches will inform our understanding of drivers behind *B. dorsalis* detections in the Los Angeles area and increase our ability to predict and prevent major future outbreaks.

Methods

Study Area & Species Data

The study uses a historical dataset of all non-native tephritid fruit fly detections in the state of California since their first detection in Hawaii over a century ago (*Bactrocera cucurbitae* in 1895; *Ceratitis capitata* in 1907).[12,14] Since the first non-native tephritid was detected in Hawaii in 1946, The CDFA has monitored a fine-scale trapping grid of roughly five baited traps per square mile of developed areas across the state. The tephritid data set includes coordinates, date of capture, life stage, sex, and species of each individual captured.[12,31,32] We conducted all analyses in R, Version 1.4.1717.[33]

The study area is generated from detection locations in the full tephritid data set for all species in all years monitored. We chose the greater Los Angeles area as our focal region for two reasons.

First, Los Angeles plays a significant role in nonnative tephritid populations at it has both many of the oldest and most recent detections. Second, trapping densities are stable within human-dominated areas – fewer traps exist in less populated regions.[7] The large, continuous sprawl of Los Angeles helps ensure the detections are the result of a similar trapping effort. We buffered all southern California detections by 4,800 meters, then selected the largest contiguous polygon, and smoothed the resulting polygon using Chaikin's corner cutting algorithm (Figure 3.1).[34–37]

We used the tephritid occurrence data for *Bactrocera dorsalis* within the study region. We applied a low level of spatial aggregation by rounding detection coordinates to three significant digits and eliminated coordinate-year duplicates. This mitigates the effect of densely spatiotemporally clustered outbreaks which can bias statistical results without providing added biological information. We selected a 10-year temporal subset of detections as our observed dataset (years 2000-2009; n=218; Figure 3.1). In future studies, we will compare the 2010-2019 occurrence data, which is currently being compiled.

The tephritid data set is presence-only: it does not provide spatiotemporal information regarding trapping locations with zero detections. Selection of pseudo-absence data for a presence-only dataset can have significant impacts on model outputs: background points with environmental data too dissimilar from those of the observations may positively bias variable parameter estimates.[38,39] Background points too similar to the observed data will fail to generate enough variation for useful inference. Numerous studies have investigated trade-offs among pseudo-absence approaches.[40,41] We used two methods to produce pseudo-absences, or background

points. First, we generated a set of spatially random points within the study area using the *sp* package (n=109).[37] Second, we used locations of all detections of non-*B. dorsalis* species (n=109). We combined these two sets of points into a single background dataset matching the abundance of the observed dataset (n=218; Figure 3.1).

Explanatory Variables

To assess the relative contributions of establishment or re-introduction in determining *B. dorsalis* occurrence, we selected explanatory variables from four categories: *B. dorsalis* population metrics, bioclimatic variables, human development and transportation metrics, and human population characteristics (Table 3.1). All variable layers were imported into R,[33] converted to the same geographic coordinate system (WGS84), and rasterized across the study area. The cells of each raster grid are approximately 1.3 km^2 in area.

To test the importance of previous *B. dorsalis* detections on future distributions, we created two distance layers of detections prior to the observed data range. The "recent" neighbors layer covered detections in the study area from year 1990 to 1999 (n=372). The "older" neighbors distance layer included year 1980 to 1989 (n=313). These layers represent the distance at any point within the study area to the nearest previous tephritid neighbor. We converted each point pattern to a distance raster where each grid cell represents the distance to a point in the dataset: a grid cell containing a detection would be zero, with the value increasing with distance from any given point.

To understand the effect of biophysical variables on *B. dorsalis* occurrences, we included tree canopy cover, elevation, and climatic layers (Table 3.1). The 2011 tree canopy data is a percent cover estimates from the U.S. Geological survey.[24,42] We obtained aggregated elevation data and the full set of 19 bioclimatic variables from WorldClim.[36,43]

To understand the effect of human development and transportation hubs, we selected candidate variables that may be important to establishment or re-introduction hypotheses (Table 3.1). Developed land cover is a pseudo-quantitative derived variable based on National Land Cover Data categories.[42,44,45] The rasterized grid cells range in value from zero to four, representing a pseudo-continuous variable where four is "developed: high density," three is "developed: medium density," two is "developed: low density," one is "developed: open space," and zero is "undeveloped", representing all other land use categories, which are primarily natural biomes. Data on airports includes the distance to any public or military airports currently permitted by the California Department of Transportation (Caltrans) Division of Aeronautics.[46,47] Bus station data represents the distance to stations in the Amtrak Thruway bus system.[48] Freight intermodal facilities are the distance to any transfer points along the California freight network for freight moving from ship to rail or truck or vice versa.[49] Port data represents distance from private or public major commercial ports as per the California Department of Commerce, Office of Economic Research.[50] Rail station data indicated distance to the nearest California passenger rail station (Table 3.1).[51]

We used multiple variables from the 2000 United States Decennial Census to explore the possible influence of human populations on the likelihood of *B. dorsalis* detections.[52] Each variable was downloaded at the census tract scale, converted to density or proportion, and rasterized.[53] The population density variable is scaled to census tract area. The remainder of the variables represent proportion of the population identifying as a single race in the following categories: white; American Indian or Alaskan Native; Black of African American; Asian; or Hawaiian or other Pacific Islander (Table 3.1).

Data analysis and modeling

We compared multiple models to identify the variables that had the strongest associations with *B. dorsalis* presences and background points. All models are global regarding the defined study area and occurrences. Multicollinearity analysis of our full variable set eliminated 20 of the 36 variables using Pearson's correlation coefficient and a cut-off of 0.6 (see Appendix).[33,54] The square root of the VIF for all final variables was between 1.00 and 2.00 (VIF cutoff values range from 3.0 to 10.0), indicating low collinearity between the remaining variables. We used boxplots paired with non-parametric Mann-Whitney U tests to analyze differences in variable distribution between presences and background points.[54]

We used random forest classification and multiple logistic regression models for the occurrence data and their corresponding explanatory variables. We implemented models in R. Random forest is a machine-learning algorithm utilizing an ensemble of classification or regression trees (CART).[30,55] In a classification setting, the output of a random forest model is the class selected by the most trees. By combining the results of many random, independently bootstrapped trees,

overall variance is decreased and overfitting is avoided.[30,56] We ran all random forest models using the *randomForest* package with 500 trees (n_{tree}).[57] Each model was individually tuned for the number of variables per tree (m_{try}). Variable importance was considered using the mean decrease in the Gini coefficient, the average total decrease of a given variable on tree node impurity, or how well the trees split the data. Models were evaluated using out-of-bag (OOB) error estimates and area under the curve (AUC) of a receiver operating characteristic (ROC) plots[58] using the *ROCR* package.[59] We ran logistic regressions with binomial distributions on standardized variables (x – mean / standard deviation).[33] We analyzed individual variables for importance using absolute value of the z statistic. We verified model assumptions using QQ plots, a Chi-Square test of residual deviance, and Pearson's residuals for the explanatory variables.[60]

Results

Descriptive Statistics

We summarized the 16 final explanatory variables by detection and background points (Table 3.2; Figure 3.2). Overall, we found little difference in the summary statistics of detections and background points for most biophysical variables. However, *B. dorsalis* occurred in areas that were significantly closer to past detections (that occurred from 1980-1989 or from 1990-1999), compared to background points (Table 3.2; Figure 3.2). Detections tended to be closer to public airports, bus stations, and rail stations, but were slightly farther from freight terminals than background points (Figure 3.1). Tree canopy cover was unexpectedly low in both categories (detection mean: 2.56%; background mean: 2.75%) compared to the maximum values (detection

max.: 35.00%; background max.: 32.00%; Table 3.2). Proportions of the population who identify as American Indian or Hawaiian represented the smallest values of all census data categories.

Importance of explanatory variables

Both *B. dorsalis* neighbor variables were included in the final model (Table 3.2). Only tree cover, annual mean temperature, minimum temperature of the coldest month, and precipitation of the wettest month met model criteria for inclusion as most bioclimatic variables are highly correlated. We found distance to public airports, bus station, freight intermodal facilities, and rail stations to be suitable variables in the development and transportation category based on collinearity. All human population metrics were included in the final model.

Overall, we found recent neighbor distance and bus station distance to be the most important explanatory variables in both models (Figure 3.3). Annual mean temperature also was highly ranked in importance in both models (3rd for random forest; 5th for logistic regression). The relative importance of all other variables was different depending on model type.

Random Forest

The random forest had an out-of-bag (OOB) error rate of 29.21% and an AUC of 0.78 (see Appendix), suggesting that the model has an acceptable rate of accuracy. Distance from a bus station was the by far the most important explanatory variable (mean decrease Gini = 28.92), followed by distance from recent, previous detections (mean decrease Gini = 18.65; Figure 3.3). Most other variables ranged in importance from 8 to 14 mean decrease Gini. Developed land cover and tree canopy were the least important variables in determining *B. dorsalis* occurrences.

Logistic Regression

As in the random forest model, distance to a recent *B. dorsalis* detection and distance to a bus terminal were the most important variables in the logistic regression model, followed by the proportion of the populations that identified as white, the minimum temperature in the coldest month, the annual mean temperature, and the proportion of the population that identified as Hawaiian (Figure 3.3; Table 3.3). As distance from a recent, previous detection point increased from the minimum observed distance (0.00 km) to the maximum (9.54 km), the probability of detecting *B. dorsalis* increased by a factor of 19.08 (Table 3.2; Table 3.3); the variable for older previous detections had a similar, albeit weaker, effect on *B. dorsalis* occurrence. *B. dorsalis* detection increased by 126.13 times at points adjacent to bus terminals, compared to points that were more than 20 km away. Detection of *B. dorsalis* increased in areas where a greater proportion of the population identified as white or as Hawaiian (Table 3.3; Figure 3.5). In addition to these anthropogenic and population-related variables, the probability of *B. dorsalis* detection significantly increased with lower minimum temperatures in the coldest months and with warmer mean temperatures (Table 3.3; Figure 3.5). The residual deviance of the logistic regression was low and not statistically significant (Chi-square, $p = 0.090$). The QQ plot and quantile regression plot showed no deviance from uniformity.

Discussion

We examined a long-term dataset of tephritid detections for evidence that repeated anthropogenic introductions are important drivers of the occurrence of a damaging, costly, and difficult-to-detect invasive pest, *B. dorsalis* fruit flies. Instead, we found evidence that *B. dorsalis* detections may be primarily driven by established populations in the Los Angeles area.

Of the 16 variables included in the final model, distance to a recent neighbor was one of the two most important variables in both models. The significance of the distance to a recent neighbor metric compared to repeated introduction variables is a strong indication of established *B. dorsalis* populations – as distance from a recent neighbor increases the probability of a future detection decreases. Older previous detections had a similar, but slightly weaker, effect on probability of *B. dorsalis* occurrence (Table 3.3; Figure 3.5), suggesting that populations may have been established for many decades.

Interestingly, tephritid detections were negatively correlated with distance to a bus station in the Amtrak throughway system, but this finding lacks a straightforward explanation (Table 3.3). The bus station data is specifically Amtrak bus stations, which connect with airports but overall are more likely to be associated with intra-state travel than international or trans-Pacific travel. Although distance to bus stations does not clearly represent an introduction pathway, it is likely correlated another variable not included in the model or the full variable list. One possibility is income, which could influence care and maintenance of neighborhood fruit trees or microclimate.[26,44,61]

Bioclimatic variables are often essential predictors of species distributions and are often a central factor in predicting invasion dynamics.[20,62,63] Though the bioclimatic variables included in the models differed in their importance between random forest and logistic regression (Figure 3.3), temperatures seem to be an important driver of *B. dorsalis* detections (Table 3.3; Figure 3.5), despite the small study area and low variation (Table 3.2). Like most other non-native tephritid

species found in California, *B. dorsalis* is a tropical species and thrives in warmer climates.[12,14] The relationships suggested by the minimum temperature of the coldest month and the precipitation of the wettest month are both counterintuitive to our knowledge of the species: as a tropical species, *B. dorsalis* thrives in humid climates and tephritids generally are thought to be sensitive to low minimum temperatures during their overwintering periods (Figure 3.4; Figure 3.5). Tree canopy cover was the least important variable in the models (Figure 3.3), though this variable did not differentiate between fruit trees and non-fruiting trees. *B. dorsalis* is a generalist species with many hosts, but fruit trees are likely a significant factor in maintaining low level populations. While neighborhoods with more tree canopy cover generally may have more fruit trees, fruit trees tend to have smaller fruit prints than larger, older ornamental species found in the greater Los Angeles area. In future analyses we hope to obtain more specific information on fruit tree density within the area.

There is little evidence from the logistic regression model that transportation metrics influence *B. dorsalis* detections, suggesting minimal importance of tephritid re-introductions through import of goods. Transportation metrics (specifically, distance to freight facilities, public airports, or rail stations) are more important variables in the random forest model than the in logistic regression model (Figure 3.3). These two statistical methods work in very different ways. Logistic regression considers how the variables work together and determines their relative effect size. In random forest, variable importance is not a measure of effect size, rather of how each variable splits the response data on its own or when interacting with another variable. Distance to a recent neighbor and distance to a bus station are the most important variables in both models, suggesting they are as important considered together as separately. Variables that are important

in the random forest model but not in the logistic regression model may be influencing the model in combination with another metric, so their overall influence is more difficult to understand (Figure 3.4). The coefficient estimate for distance to freight facilities was positive, so if the effect had been significant, further distance from freight intermodal facilities would increase the probability of detecting *B. dorsalis*.

Our results show little support for the hypotheses of re-introduction by travel. Neither distance to public airports nor all human population metrics were consistently important between the two model types (Figure 3.3). Proportion of the population identifying as Hawaiian was a significant explanatory variable in the logistic regression model, but white and Black were as well (Table 3.3; Figure 3.5). However, we hypothesize that these demographic variables may be correlated with other underlying anthropogenic drivers of *B. dorsalis* occurrence. Census-related variables may be picking up on signals from other variables such as income, host plant distribution, or a yet unidentified variable. It is possible that the census data, or other variables, does not capture the scale relevant to micro populations. The study area was designed to minimize certain variables, such as land cover type and climate. However, gradients biologically significant to an individual fly might be the those captures by census tracts or even a ~2km raster grid cell. Some metrics, like census data, may change block to block, even house to house, while flies have potential dispersal distances of miles.[64]

There are several anthropogenic variables not included in this study that may improve future models. City metrics such as neighborhood age or income could be important in understanding

neighborhoods that are more likely to have fruit trees that perhaps are less carefully managed. Another avenue is a deeper look into historical data and the corresponding explanatory variables. For example, early investigations indicated high correlation (Pearson's correlation coefficient <0.8) between distance to military airports and detections before 1975. It has been suggested that early tephritid introductions were made as far back as World War II and the Vietnam war when soldiers were traveling back from areas with *Bactrocera* spp. tephritids.[65] Though difficult, retrieving historical population and urbanization data could prove a fruitful avenue of investigation.

As the proportion of developed land increases, invasion research must consider human-dominated landscapes.[18,66–68] Effective modeling species in human-dominated landscapes may include many of the same predictors as species in natural areas, but many neglect to incorporate some human-related variables.[61] Inclusions of metrics such as human population density, income, and neighborhood age may be critical components to understanding species dynamics in human-dominated landscapes. We encourage the use of this vast trove of spatial information in combination with more traditional environmental metrics when relevant. Understanding these links will only prove more important in understanding future species invasions.

Acknowledgments

We thank Dr. Allison Simler for guidance on approach, modeling, and the manuscript. We also thank Dr. Jay Rosenheim and Dr. Jenny VanWyk for helpful feedback on the manuscript. We thank the Carey and Hijmans lab for advice. This material is based on work supported by National Science Foundation Graduate Research Fellowship Grant. Any opinions, findings, and

conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation. Specific author contributions: CCL conceived the study, analyzed data, and wrote the initial draft. RJH assisted with coding and analysis. JRC provided the dataset, expertise on the study species, and contributed substantial feedback at each stage of the project.

Works Cited

1. Gormley, A. M. *et al.* Using presence-only and presence-absence data to estimate the current and potential distributions of established invasive species. *J. Appl. Ecol.* **48**, 25–34 (2011).
2. Davis, M. A. *Invasion biology*. (Oxford University Press, 2009).
3. Triplehorn, C. A., Johnson, N. F. & Borror, D. J. *Borror and DeLong's Introduction to the Study of Insects*. **7**, (Thomson Brooks/Cole, 2005).
4. Marshall, S. A. *Flies: The Natural History & Diversity of Diptera*. (Firefly, 2012).
5. Liquido, N. J. *et al.* A review of recorded host plants of Oriental Fruit Fly, Bactrocera (Bactrocera)dorsalis(Hendel)(Diptera: Tephritidae), version 3.0. *USDA CPHST Online Database* (2017).
6. California Department of Food & Agriculture & U.S. Department of Agriculture. Exotic Fruit Fly Regulatory Resonse Manual. (2001). doi:10.1093/infdis/jit776
7. Gilbert, A. J., Bingham, R. R., Nicolas, M. A. & Clark, R. A. *Insect Trapping Guide*. (2013).
8. California Department of Food & Agriculture. Oriental Fruit Fly Fact Sheet. (2018). Available at: https://www.cdfa.ca.gov/plant/factsheets/OFF_FactSheet.pdf.
9. *Medfly infestation triggers quarantine in central Los Angeles*. (2014).
10. California Department of Food & Agriculture. *California Agricultural Statistics Review*. (2020).
11. Bateman, M. The ecology of fruit flies. *Annu. Rev. Entomol.* **17**, (1972).
12. Papadopoulos, N. T., Plant, R. E. & Carey, J. R. From trickle to flood: the large-scale, cryptic invasion of California by tropical fruit flies. *Proc. Biol. Sci.* **280**, 20131466 (2013).
13. Aluja, M. & Norrbom, A. *Fruit Flies (Tephritidae): Phylogeny and Evolution of Behavior*. (CRC Press, 2010).
14. Stephens, A. E. A., Kriticos, D. J. & Leriche, A. The current and future potential geographical distribution of the oriental fruit fly, Bactrocera dorsalis (Diptera: Tephritidae). *Bull. Entomol. Res.* **97**, 369–378 (2007).
15. Papadopoulos, N. T., Plant, R. E. & Carey, J. R. From trickle to flood : the large-scale , cryptic invasion of California by tropical fruit flies. *Proc. R. Soc.* (2013).
16. Zhao, Z. *et al.* Life table invasion models: spatial progression and species-specific partitioning. *Ecology* **100**, 1–11 (2019).
17. Liebhold, A. M. & Tobin, P. C. Population ecology of insect invasions and their management. *Annu. Rev. Entomol.* **53**, 387–408 (2008).
18. Davis, M. A. *Invasion Biology*. (Oxford University Press, 2009).
19. Carroll, S. P. Conciliation biology: the eco-evolutionary management of permanently invaded biotic systems. *Evol. Appl.* **4**, 184–199 (2011).
20. Gaston, K. J. *The Structure and Dynamics of Geographic Ranges*. *Oxford Series in Ecology and Evolution* (2003). doi:10.2167/jost191b.0

21. Peterson, A. T. *et al. Ecological Niches and Geographic Distributions*. (Princeton University Press, 2011).
22. Bale, J. S. & Hayward, S. a L. Insect overwintering in a changing climate. *J. Exp. Biol.* **213**, 980–94 (2010).
23. Taha, H. Characterization of urban heat and exacerbation: Development of a heat Island index for California. *Climate* **5**, 18–20 (2017).
24. Coulston, J. W. *et al.* Modeling percent tree canopy cover: A pilot study. *Photogramm. Eng. Remote Sensing* **78**, 715–727 (2012).
25. Grosberg, R. K., Vermeij, G. J. & Wainwright, P. C. Biodiversity in water and on land. *Curr. Biol.* **22**, R900-3 (2012).
26. Hall, S. J. *et al.* Convergence of microclimate in residential landscapes across diverse cities in the United States. *Landsc. Ecol.* **31**, 101–117 (2016).
27. Klaus I, S., James R, S. & E Gregory, M. Effetcs of tree cover on parking lot microclimate and vehicle emissions. *J. Arboric.* **25**, 129–142 (1999).
28. Mi, C., Huettmann, F., Guo, Y., Han, X. & Wen, L. Why choose Random Forest to predict rare species distribution with few samples in large undersampled areas? Three Asian crane species models provide supporting evidence. *PeerJ* (2017). doi:10.7717/peerj.2849
29. Chakraborty, A., Gelfand, A. E., Wilson, A. M., Latimer, A. M. & Silander, J. A. Point pattern modelling for degraded presence-only data over large regions. *J. R. Stat. Soc. Ser. C Appl. Stat.* **60**, 757–776 (2011).
30. Breiman, L. Random forests. *Mach. Learn.* **45**, 5–32 (2001).
31. Carey, J. R. Establishment of the Mediterranean Fruit Fly in California. *Science (80-.).* **253**, 1369–1373 (1991).
32. Carey, J. R. The future of the Mediterranean fruit fly Ceratitis capitata invasion of California: A predictive framework. *Biol. Conserv.* **78**, 35–50 (1996).
33. R Core Team. R: A language and environment for statistical computing. (2021).
34. Strimas-Mackey, M. smoothr: Smooth and Tidy Spatial Features. (2021).
35. Bivand, R. S. & Rundel, C. rgeos: Interface to Geometry Engine - Open Source ('GEOS'). (2020).
36. Hijmans, R. J. raster: Geographic Data Analysis and Modeling. (2021).
37. Pebesma, E. & Bivand, R. Classes and methods for spatial data in R. *R News* **5**, (2005).
38. Hazen, E. L. *et al.* Where did they not go? Considerations for generating pseudo-absences for telemetry-based habitat models. *Mov. Ecol.* **9**, 1–13 (2021).
39. Valavi, R., Elith, J., Lahoz-Monfort, J. J. & Guillera-Arroita, G. Modelling species presence-only data with random forests. *bioRxiv* 1–18 (2020). doi:10.1101/2020.11.16.384164
40. Engler, R., Guisan, A. & Rechsteiner, L. An improved approach for predicting the distribution of rare and endangered species from occurrence and pseudo-absence data. *J. Appl. Ecol.* **41**, 263–274 (2004).

41. Guisan, A. & Zimmermann, N. E. Predictive habitat distribution models in ecology. *Ecol. Modell.* **135**, 147–186 (2000).
42. Bocinsky, R. K. FedData: Functions to Automate Downloading Geospatial Data Available from Several Federated Data Sources. (2020).
43. Fick, S. E. & Hijmans, R. J. WorldClim 2: new 1-km spatial resolution climate surfaces for global land areas. *Int. J. Climatol.* **37**, 4302–4315 (2017).
44. Homer, C. *et al.* Conterminous United States land cover change patterns 2001–2016 from the 2016 National Land Cover Database. *ISPRS J. Photogramm. Remote Sens.* **162**, 184–199 (2020).
45. Jin, S. *et al.* Overall methodology design for the United States national land cover database 2016 products. *Remote Sens.* **11**, (2019).
46. State of California. Public Airports. *California State Geoportal* (2019).
47. State of California. Military Airports. *California State Geoportal* (2019). Available at: https://gis.data.ca.gov/datasets/ffd05c8c5e0047fea35390864810da31_0/explore?location=36.406741%2C-118.860000%2C6.49. (Accessed: 2nd June 2021)
48. State of California. Amktrak Bus Stations. *California State Geoportal* (2019).
49. State of California. Freight Intermodal Facilities. *California State Geoportal* (2019). Available at: https://gis.data.ca.gov/datasets/f6bab2fbee864879896a9f4a881b84b0_0/about. (Accessed: 2nd June 2021)
50. State of California. Ports. *California State Geoportal* (2020).
51. State of California. California Rail Stations. *California State Geoportal* (2019).
52. U.S. Census Bureau. *Decennial Census Summary File 1*. (2000).
53. Walker, K. & Herman, M. idycensus: Load US Census Boundary and Attribute Data as 'tidyverse' and 'sf'-Ready Data Frames. (2021).
54. Harrell Jr, F. E. c: Harrell Miscellaneous. (2021).
55. Hastie, T., Tibshirani, R. & Friedman, J. Random Forest. in *The Elements of Statistical Learning* **27**, 83–85 (2009).
56. Elith, J. Chapter 6 : Predicting distributions of invasive species. (2011).
57. Liaw, A. & Wiener, M. Classification and Regression by randomForest. *R News* **2**, 18–22 (2002).
58. Hanley, J. A. & McNeil, B. J. The Meaning and Use of the Area under Receiver Operating Characteristic (ROC) Curve. *Radiology* **143**, 29–36 (1982).
59. Sing, T., Sander, O., Beerenwinkel, N. & Lengauer, T. ROCR: visualizing classifier performance in R. *Bioinformatics* **21**, (2005).
60. Hartig, F. DHARMa: Residual Diagnostics for Hierarchical (Multi-Level / Mixed) Regression Models. (2021).
61. Loss, S. R., Ruiz, M. O. & Brawn, J. D. Relationships between avian diversity, neighborhood age, income, and environmental characteristics of an urban landscape. *Biol. Conserv.* **142**, 2578–2585 (2009).

62. Robinet, C. & Roques, A. Direct impacts of recent climate warming on insect populations. *Integr. Zool.* **5**, 132–42 (2010).
63. Biber-Freudenberger, L., Ziemacki, J., Tonnang, H. E. Z. & Borgemeister, C. Future risks of pest species under changing climatic conditions. *PLoS One* **11**, 1–17 (2016).
64. Froerer, K. M. *et al.* Long-distance movement of Bactrocera dorsalis (Diptera: Tephritidae) in Puna, Hawaii: How far can they go? *Am. Entomol.* **56**, 88–95 (2010).
65. Fullaway, D. The oriental fruit fly in Hawaii. *Proc. Entomol. Soc. Washingt.* **51**, 181–205 (1953).
66. Walther, G.-R. *et al.* Alien species in a warmer world: risks and opportunities. *Trends Ecol. Evol.* **24**, 686–93 (2009).
67. Vitousek, P. M., D'Antonio, C. M., Loope, L. L., Rejmanek, M. & Westbrooks, R. Introduced species: A significant component of human-caused global change. *N. Z. J. Ecol.* **21**, 1–16 (1997).
68. *Encyclopedia of Biological Invasions*. (University of California Press, 2011).
69. Wickham, H. ggplot2: Elegant Graphics for Data Analysis. (2016).

Figures

Table 3.1.

Category	Variable	Description	Years	Sources
Tephritid neighbor metrics	Recent neighbor distance	Distance to nearest recent prior B. dorsalis detection.	1990-1999	CDFA
	Old neighbor distance	Distance to nearest B. dorsalis detection prior to 1975.	1980-1989	CDFA
Biogeo-climatic	Tree canopy	30-meter raster geospatial dataset containing percent tree canopy estimates for each pixel across all land covers and types.	2011	USFS[24]
	Elevation	Data were aggregated from SRTM 90-meter resolution data		WorldClim[43]
	Climatic variables	0.5 minute resolution ・Annual mean temperature ・Mean diurnal range (max temp - min temp) ・Isothermality (annual mean temp / annual range) (×100) ・Temperature annual range ・Temperature seasonality (standard deviation ×100) ・Max temperature of warmest month ・Min temperature of coldest month ・Temperature annual range ・Mean temperature of wettest quarter ・Mean temperature of driest quarter ・Mean temperature of warmest quarter	2020	WorldClim[43]

		• Mean temperature of coldest quarter • Annual precipitation • Precipitation of wettest month • Precipitation of driest month • Precipitation seasonality (coefficient of variation) • Precipitation of wettest quarter • Precipitation of driest quarter • Precipitation of warmest quarter • Precipitation of coldest quarter		
Human transportation and development	Developed land cover	National Land Cover Data. 20 original classifications. 4 classifications in the "developed" category converted to pseudo-continuous values. "Developed, open space" = 1; "Developed, low intensity" = 2; "Developed, medium intensity" = 3; "Developed, high intensity" = 4. All other categories = 0.	2001	USGS[44]
	Public airports distance	Distance to the nearest public airports currently permitted by Caltrans Division of Aeronautics. Original point layer assembled by Caltrans, Division of Research, Innovation and System Information, GIS Branch.	Updated 2020	CA State Geoportal[46]
	Military airports distance	Distance to military airports currently permitted by the State of California, Department of Transportation (Caltrans), Division of Aeronautics.	Updated 2020	CA State Geoportal[47]
	Bus station distance	Distance to the nearest bus station in the Amtrak Thruway Bus system.	Updated 2020	CA State Geoportal[48]
	Freight intermodal facilities distance	Distance to intermodal freight facility terminals (transfer points to move freight from ship to rail or truck) on the California freight network.	Updated 2020	CA State Geoportal[49]
	Port distance	Distance to the nearest public or private ports as defined by California's Major Commercial Ports -1986, by the California Department of Commerce, Office of Economic Research.	Updated 2020	CA State Geoportal[50]

	Rail station distance	Distance to the nearest California passenger rail station compiled from Amtrak Operating Timetable 45 and commuter rail websites for Metrolink, ACE, Caltrain, and Coaster"	Updated 2020	CA State Geoportal[51]
US census	Population density	Total population size per census tract area.	2000	U.S. Census[52]
	Proportion of the population white	Proportion of the population of one race identifying as one race, white alone.	2000	U.S. Census[52]
	Proportion of the population American Indian or Alaskan Native	Proportion of the population of one race identifying as one race, American Indian or Alaskan Native alone.	2000	U.S. Census[52]
	Proportion of the population Black	Proportion of the population of one race identifying as one race, Black of African American alone.	2000	U.S. Census[52]
	Proportion of the population Asian	Proportion of the population of one race identifying as one race, Asian alone.	2000	U.S. Census[52]
	Proportion Hawaiian or other Pacific Islander	Proportion of the population of one race identifying as one race, Native Hawaiian and other Pacific Islander alone.	2000	U.S. Census[52]

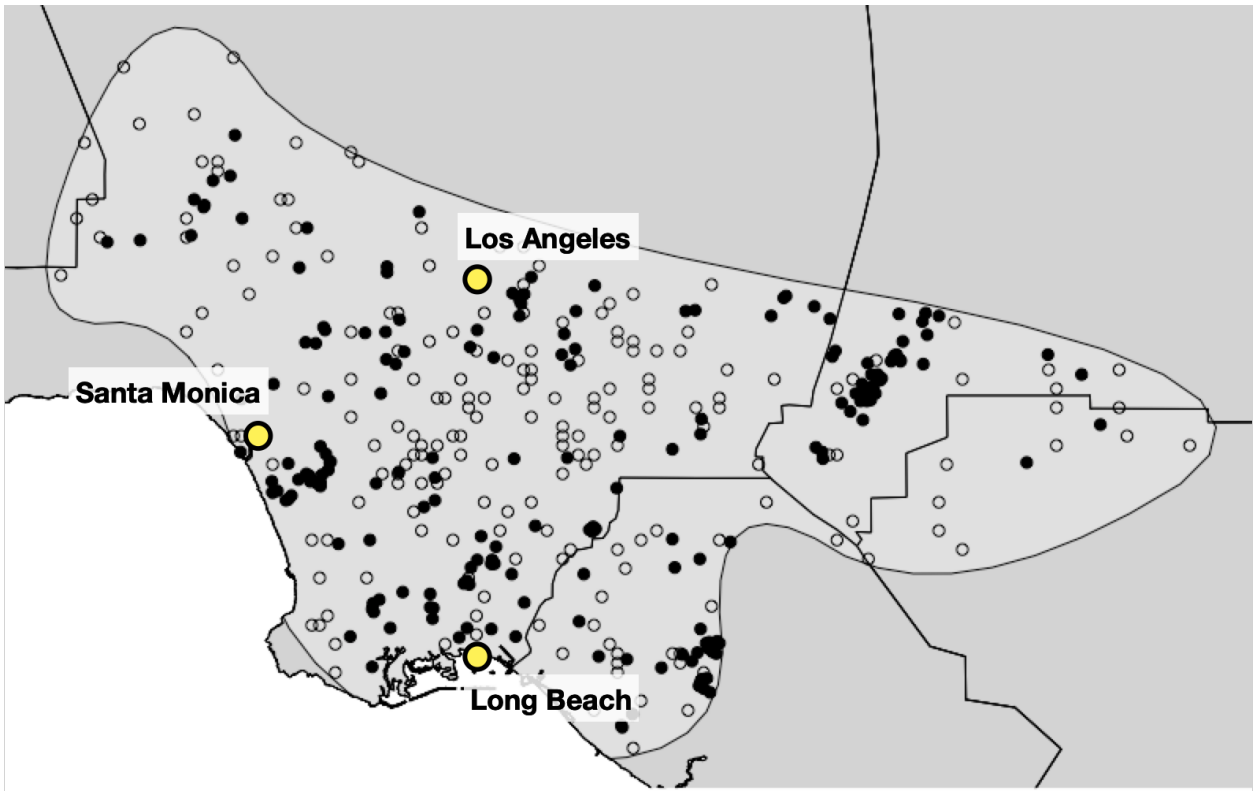
List of all explanatory variables considered pre-multicollinearity analysis.

Table 3.2.

Variable		Detections				Background			
Name	Units	Min	Max	Mean	SD	Min	Max	Mean	SD
Recent neighbor distance	kilometer	0	9.54	2.09	1.79	0	15.65	3.76	2.80
Older neighbor distance	kilometer	0	13.56	2.85	2.16	0	22.53	3.94	3.68
Tree canopy	percent cover	0	35.00	2.56	5.38	0	32.00	2.75	6.19
Annual mean temp.	°C	16.54	18.30	17.81	0.29	14.74	18.30	17.72	0.58
Min temp. coldest month	°C	3.64	9.00	6.57	1.32	2.45	8.98	6.40	*1.40*
Precip. of wettest month	millimeter	43.97	118.63	77.85	11.66	37.30	117.89	81.72	14.38
Developed land cover	ordered category	0	4.00			0	4.00		
Public airport distance	kilometer	0	21.00	7.83	4.97	0	23.03	8.91	4.52
Bus station distance	kilometer	0.95	16.96	6.09	3.74	0	21.49	9.68	4.87
Freight distance	kilometer	0	55.21	20.31	10.70	1.68	65.04	17.56	13.66
Rail station distance	kilometer	0.95	13.44	4.64	2.74	0.00	16.25	5.70	2.97
Population density	people per km^2	134.24	13485.53	3676.65	2431.07	0.11	14785.51	3220.04	2849.10
Prop. White	proportion of total	0.05	0.90	0.56	0.20	0.04	1.00	0.55	0.21
Prop. American Indian	proportion of total	0	0.02	0.01	0	0	0.02	0.01	0.01
Prop. Black	proportion of total	0	0.88	0.08	0.14	0	0.88	0.07	0.12
Prop. Asian	proportion of total	0	0.63	0.12	0.14	0	0.65	0.14	0.16
Prop. Hawaiian	proportion of total	0	0.06	0	0.01	0	0.03	0.00	0

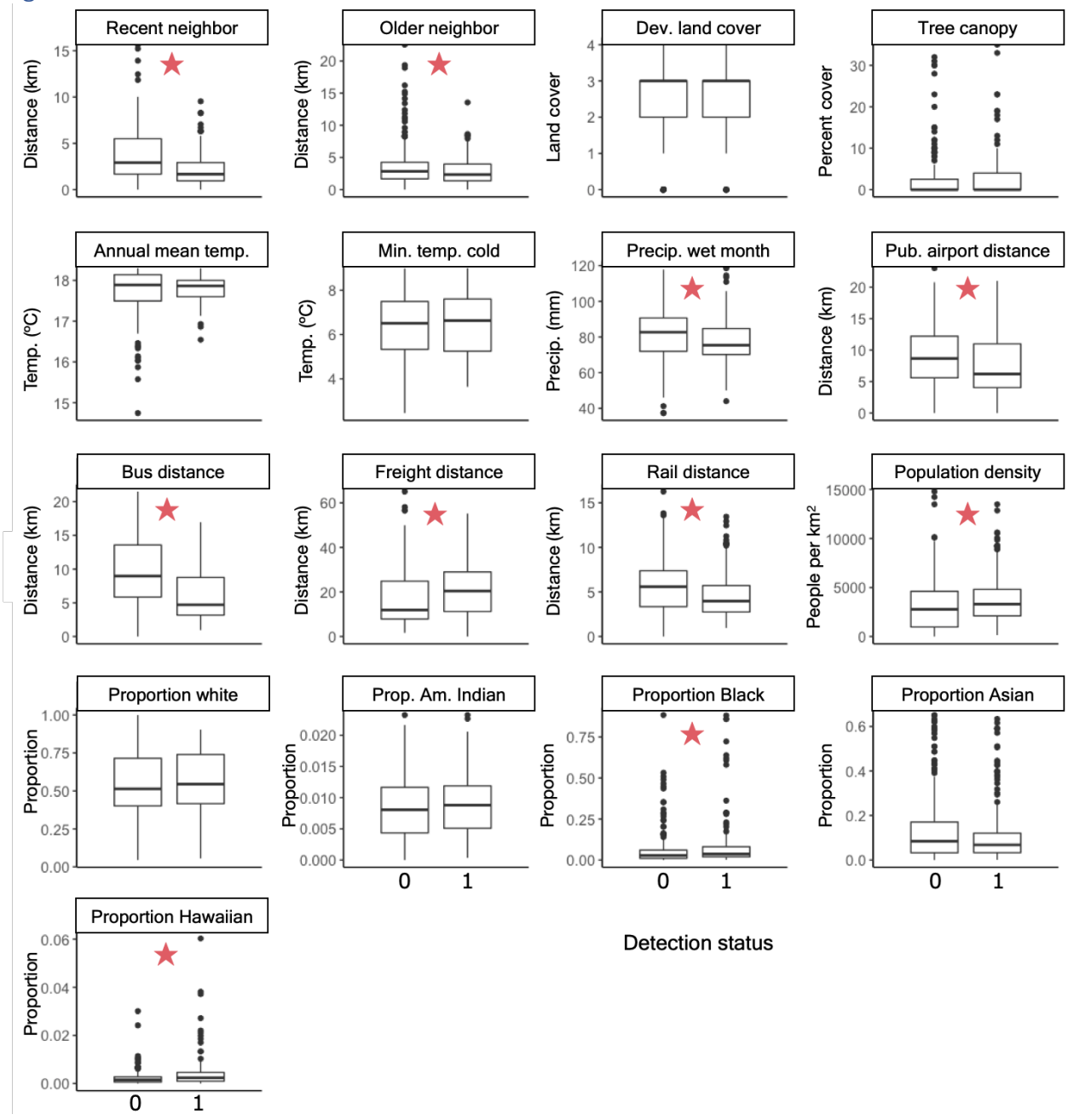
Summary statistics (the minimum, maximum, mean, and standard deviation) for final model variables for *B. dorsalis* detections and background points. The developed land cover variable contains empty cells due to being a pseudo-continuous, but still categorical, variable. Black horizontal lines in the table denote the four categories of variables: *B. dorsalis* population metrics, bioclimatic variables, human development and transportation metrics, and human population characteristics.

Figure 3.1.



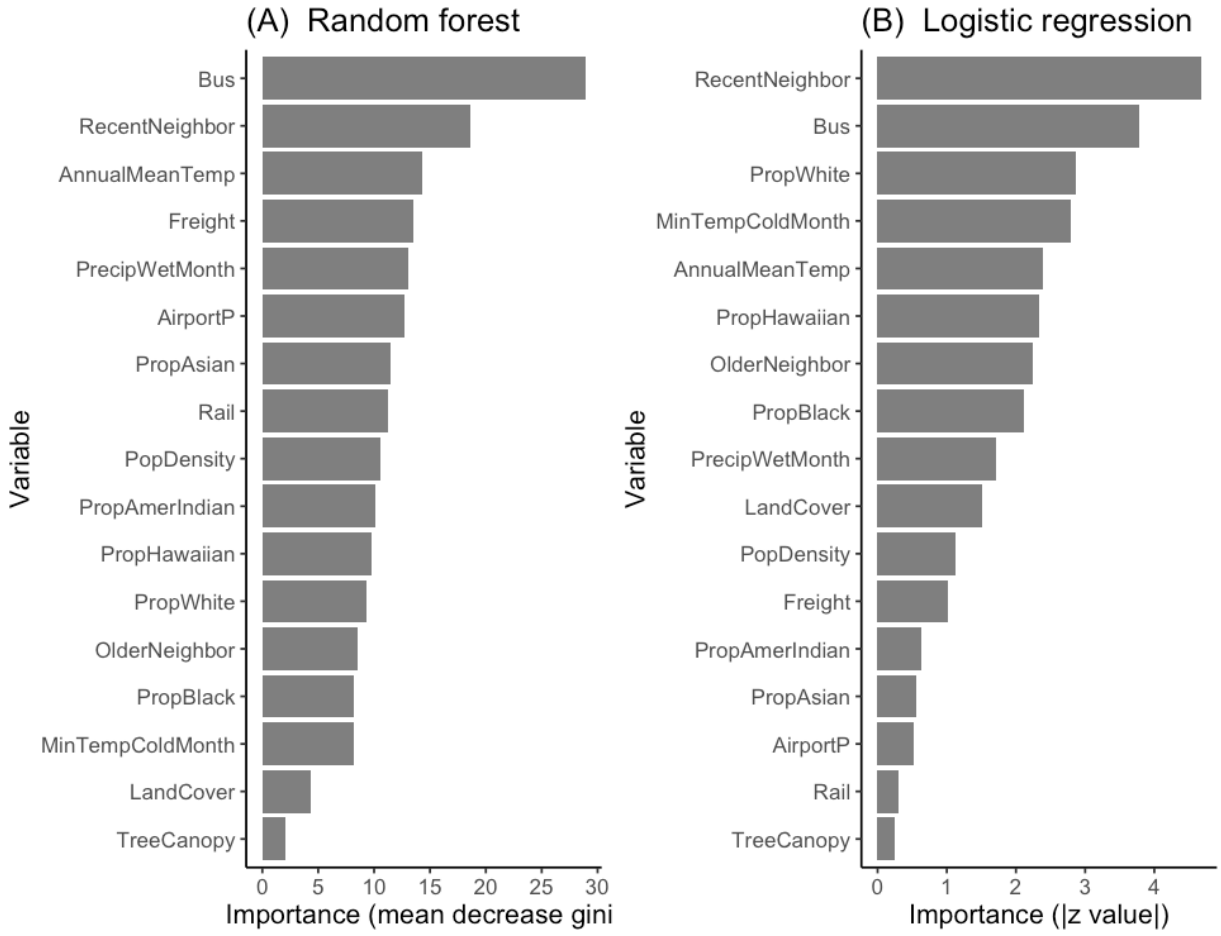
Map of the study area in Los Angeles, California. Light grey polygon represents the study area boundary. Black points represent detections of *B. dorsalis* between 2000 and 2009. Empty circles represent the pseudo-absence background points.

Figure 3.2.



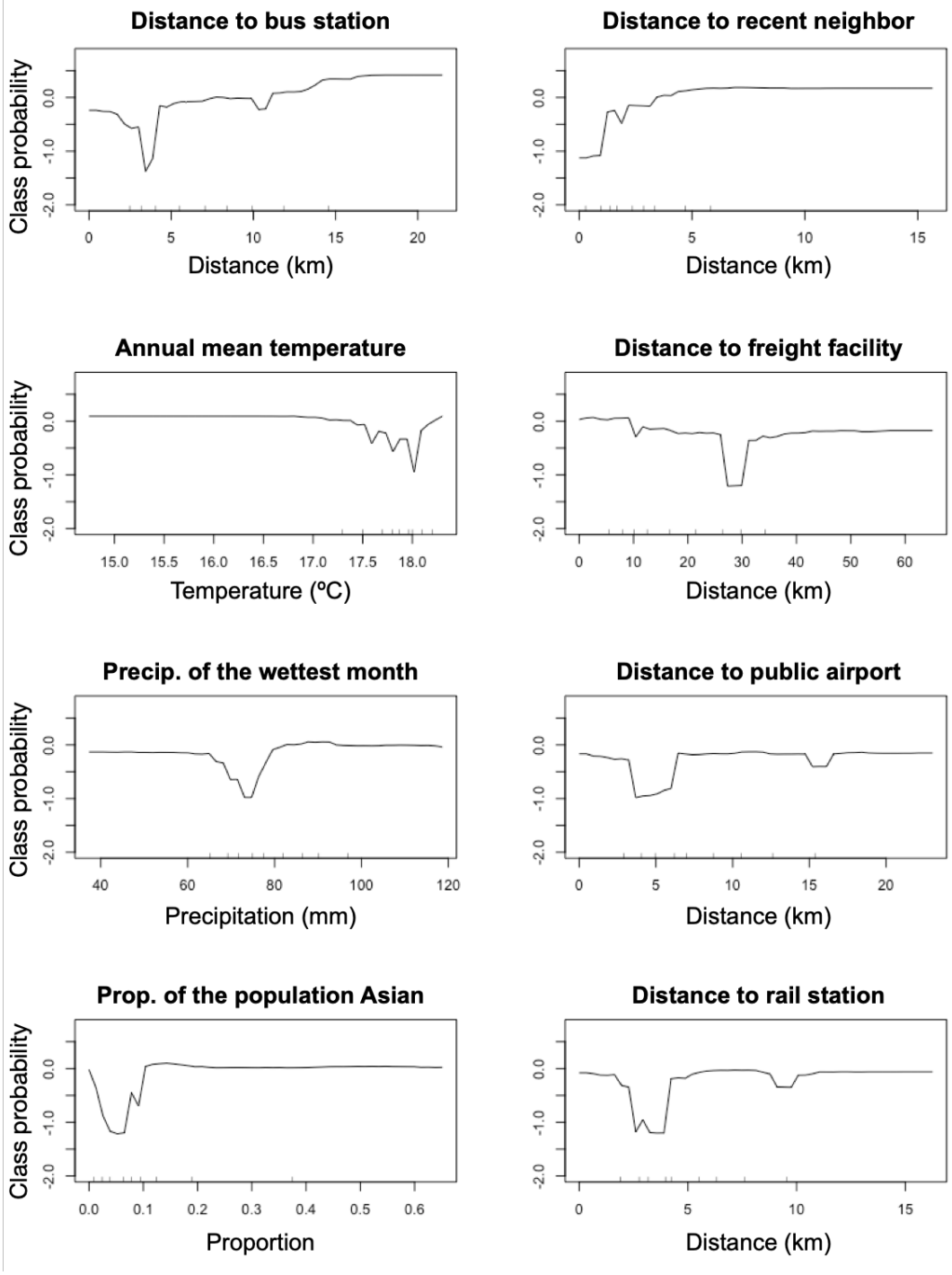
Boxplots summarize the distribution of each explanatory variable by background point, 0, or presence, 1. Red stars represent statistical differences between variables values for presences and pseudo-absences at the 5% significance level using a Mann-Whitney U test.[54,69]

Figure 3.3.



Variable importance for random forest (A) and logistic regression (B) models.

Figure 3.4.



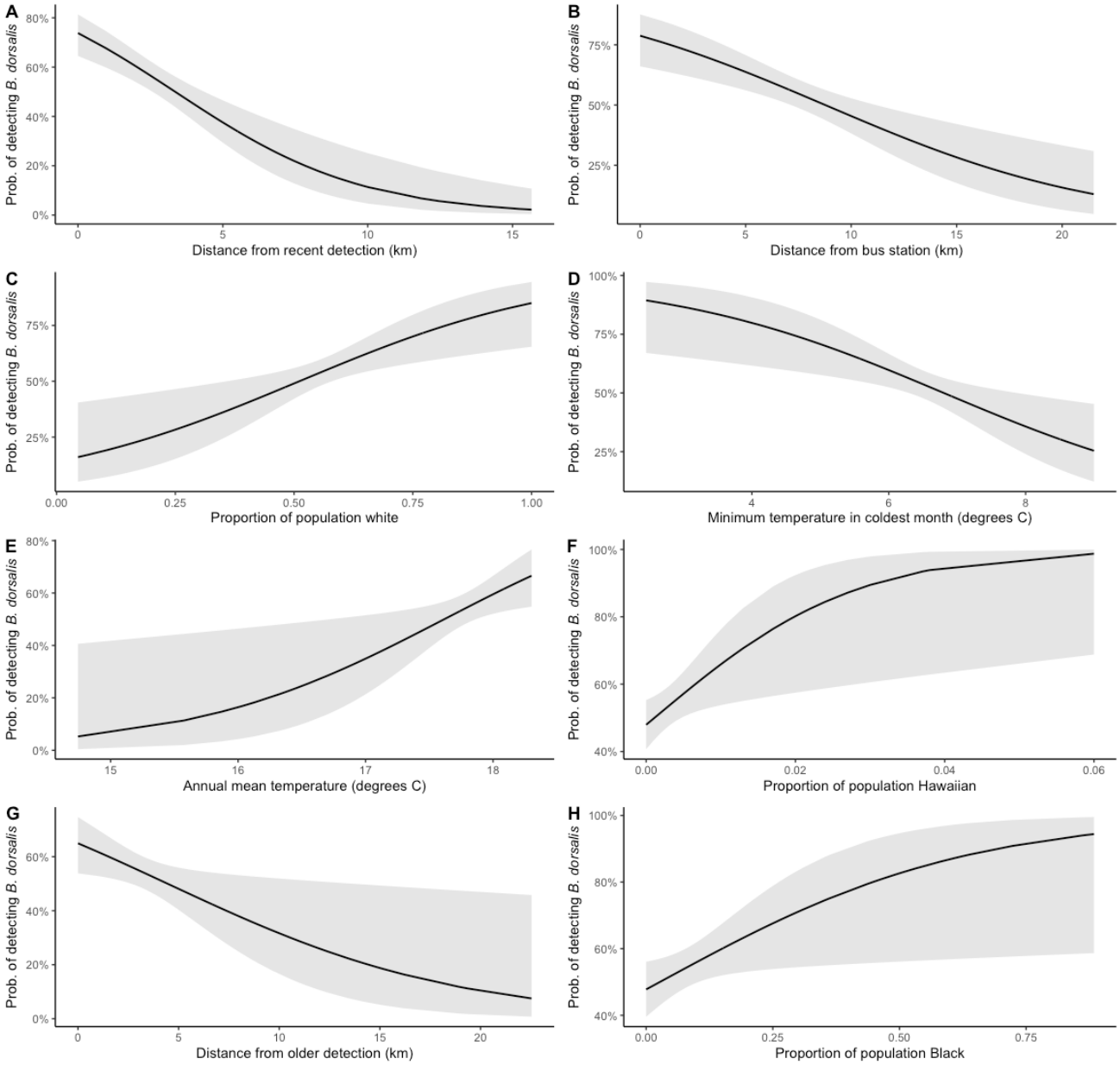
Partial dependence plots of top eight most important explanatory variables from the Random Forest model. The sharp negative spikes are the result of skew from a single dense spatiotemporal outbreak.[57]

Table 3.3.

Variable	Estimate	z value	Pr(>|z|)		Odds ratio (95% CI)
(Intercept)	0.15263	1.26	0.20769		1.16 (0.92, 1.48)
Recent neighbor distance	-0.75869	-4.672	2.99E-06	***	0.47 (0.34, 0.64)
Bus station distance	-0.69469	-3.779	0.000158	***	0.50 (0.34, 0.71)
Prop. White	0.71699	2.869	0.004122	**	2.05 (1.27, 3.38)
Min temp. coldest month	-0.66644	-2.786	0.00534	**	0.51 (0.32, 0.82)
Annual mean temp.	0.45456	2.391	0.016813	*	1.58 (1.09, 2.31)
Prop. Hawaiian	0.38821	2.332	0.01972	*	1.47 (1.11, 2.13)
Older neighbor distance	-0.41739	-2.248	0.024581	*	0.66 (0.45, 0.94)
Prop. Black	0.42695	2.118	0.034178	*	1.53 (1.04, 2.30)
Precip. wet month	-0.31732	-1.716	0.086085	.	0.73 (0.51, 1.05)
Developed land cover	0.2274	1.503	0.132797		1.26 (0.93, 1.69)
Population density	0.18904	1.117	0.263906		1.21 (0.87, 1.70)
Freight distance	0.22557	1.016	0.309506		1.25 (0.81, 1.94)
Prop. American Indian	-0.12146	-0.633	0.526978		0.89 (0.61, 1.29)
Prop. Asian	0.10595	0.555	0.579126		1.11 (0.77, 1.62)
Public airport distance	-0.07173	-0.521	0.602522		0.93 (0.71, 1.22)
Rail station distance	0.05025	0.311	0.756068		1.05 (0.77, 1.45)
Tree canopy	-0.03149	-0.239	0.811402		0.97 (0.75, 1.26)

Logistic regression coefficient estimates. Asterisks represent significance: '.' = < 0.1; '*' = < 0.05; '**' = < 0.001; '***' = 0.001-0. Positive coefficient estimates represent a positive relationship with occurrence likelihood; negative coefficient estimates indicate a negative relationship with occurrence likelihood. All variables are standardized (x – mean/standard deviation).

Figure 3.5.



Marginal effects of select variables from binomial generalized linear mixed model of *B. dorsalis* occurrence in the Los Angeles area. Effects shown were calculated while holding all other variables in the model at their mean value and include: A) the distance from previous, recent (1990-1999) detections; B) Distance from an Amtrak bus station; C) Proportion of the population

that identifies as white; D) Minimum temperature in the coldest month; E) Annual mean temperature; F) The proportion of population that identifies as Hawaiian; G) Distance from older, previous (1980-1989) detections; and H) The proportion of population that identifies as Black. Black lines indicate the mean predicted trend, with 95% confidence intervals displayed in grey.

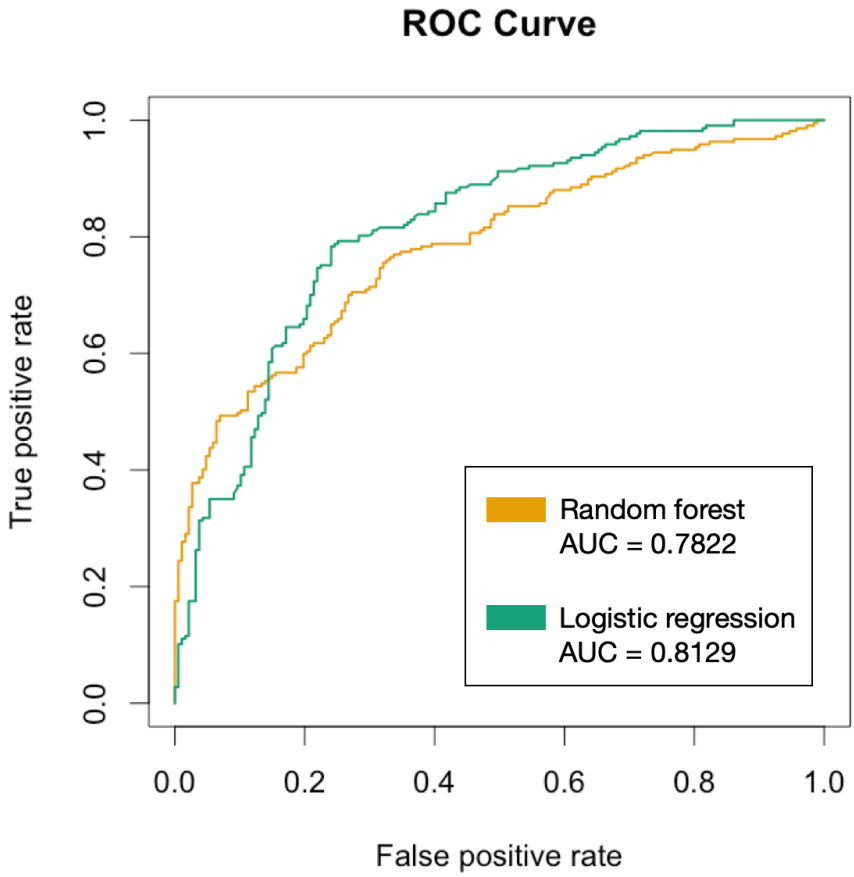
Appendix

Table 3.1.

Variable	Recent neighbor distance	Older neighbor distance	Developed land cover	Tree canopy	Annual mean temp.	Min temp. of coldest month	Precip. of wettest month	Public airport distance	Bus station distance	Freight distance	Rail station distance	Population density	Prop. White	Prop. American Indian	Prop. Black	Prop. Asian	Prop. Hawaiian
Recent neighbor distance	-																
Older neighbor distance	0.49	-															
Developed land cover	-0.3	-0.29	-														
Tree canopy	-0.06	-0.03	-0.26	-													
Annual mean temp.	-0.36	-0.33	0.31	-0.04	-												
Min temp. coldest month	-0.34	-0.47	0.44	-0.08	0.33	-											
Precip. wettest month	0.05	0.04	-0.17	0.27	-0.2	-0.33	-										
Public airport distance	0.12	-0.09	-0.11	0	-0.27	-0.07	0.08	-									
Bus station distance	0.34	0.29	-0.07	-0.06	0.06	-0.1	0.02	0.04	-								
Freight distance	0.27	0.37	-0.33	-0.1	-0.43	-0.56	-0.22	0.09	-0.23	-							
Rail station distance	0.19	0.12	-0.06	-0.03	-0.22	0.19	-0.1	0.09	0.46	-0.07	-						
Population density	-0.24	-0.37	0.38	-0.1	0.24	0.44	-0.14	0.21	-0.21	-0.28	-0.11	-					
Prop. White	0.23	0.36	-0.2	0.13	-0.43	-0.25	0.09	0.06	0.03	0.3	0.08	-0.38	-				
Prop. American Indian	-0.08	-0.17	0.08	-0.27	0.34	-0.06	-0.27	0.03	-0.04	0.11	-0.25	0.23	-0.29	-			
Prop. Black	-0.07	-0.04	0.04	-0.1	0.12	0.28	-0.08	-0.21	0	-0.1	0.22	0.05	-0.48	-0.1	-		
Prop. Asian	-0.01	-0.07	-0.02	0.15	-0.02	-0.01	0.18	0.05	0.09	-0.24	0.04	-0.12	-0.25	-0.41	-0.17	-	
Prop. Hawaiian	-0.02	-0.09	0.14	-0.09	0.03	0.23	-0.25	-0.14	-0.11	-0.13	0.03	0.11	-0.21	0	0.13	0.09	-

Pearson's correlation coefficient matrix among all explanatory variables of the final models.

Figure 3.1.



Receiver operating characteristic (ROC) plots for each model. Area under the ROC curve (AUC) values are included in the legend. The closer a given ROC curve is to the upper left corner, the better the model performance with zero false positives (100% specificity) and only true positives (100% sensitivity). The dashed grey line represents the "line of no discrimination": the performance of a random guess classifier (AUC=0.5).[59]