*Landscape-scale urban pollinator occupancy rates are associated with natural habitat area and income.*

**Abstract**

As urbanization expands globally, there is a need to better understand the major drivers of urban biodiversity dynamics and the spatial and temporal scales at which these drivers operate. This need is critical for functionally important and widely declining insect pollinators. Although previous studies have determined that natural habitat and floral resources drive locally observed patterns of urban pollinator biodiversity, we lack an understanding of whether these drivers influence urban pollinator biodiversity patterns at the landscape-scale. To test the hypotheses that greater total area of undeveloped natural habitat and greater average income (with income used as a proxy for plant diversity and management resources) in an urban landscape promote pollinator biodiversity at the landscape scale, we applied an integrated Bayesian occupancy model to 12 years of natural history collections records, comparing occurrence rates of bumble bee and hoverfly pollinator species across 459 urban landscapes from across the United States. We found that landscape-scale occupancy rates for the average bumble bee species is not significantly associated with the area of natural habitat in an urban landscape (95% BCI for effect of an increase in 1 standard deviation in total natural habitat area = – ), however, we found a positive association with average income (50% BCI = –). In contrast, we found a positive association between natural habitat area and hoverfly occupancy rate (95% BCI = ), emphasizing that natural habitat remnants such as greenbelts and nature reserves are critical for the maintenance of pollinator diversity in urban landscapes. Together, these landscape-scale results suggest that local pollinator population dynamics in urban areas are driven in part by processes that operate at spatial scales beyond those experienced by individual organisms. Given these results, local urban habitat restoration and enhancements must coordinate with city and regional planning policies to ensure the long-term success of species conservation.

1. ***Introduction***

The intensity and extent of urban land use is increasing globally (Seto et al., 2012; Simkin et al., 2022). By 2050, over 70% of the world’s population is projected to live in urban areas (United Nations, 2018). At the same time, urbanization is associated with changes to the spatio-temporal composition and configuration of critical habitat and other resources, with the potential for negative impacts on biodiversity (McKinney, 2006; Grimm et al., 2008; Kremen & Jha 2013b; Leong et al., 2016; Eichenberg et al. 2021). In order to reduce the loss of biodiversity and associated ecosystem services in an urbanizing world, we need to better understand the drivers of urban biodiversity patterns and the spatial and temporal scales at which they operate. Identifying these drivers as well as their scale dependence is especially pressing for pollinators, a widely threatened functional group (Biesmeijer et al., 2006; Potts et al., 2010; Koh et al. 2016) that provides valuable ecosystem services across a variety of environments, including in urban systems (pollination of native plants, e.g., Cheptou et al, 2006; and pollination for urban agriculture, e.g., Potter et al., 2015, Lin et al., 2015; Jha et al., 2023).

Broadly, insect pollinator declines are driven by a combination of pressures including climate change (Oliver et al., 2017; Jackson et al., 2022; Janousek et al., 2023), species invasions (Valtonen et al. 2006; Moron et al., 2009; Kaiser-Bunbury et al. 2017, Aizen et al., 2019), pesticide use (Brittain et al., 2010; Rundlof et al., 2015; Janousek et al., 2023) and disease (Colla et al., 2006; Furst et al., 2014). Anthropogenic land use change may exacerbate insect pollinator declines through habitat loss and habitat fragmentation and, further, by amplifying other direct stressors (Winfree et al., 2009). For example, agricultural land simplification often drives greater pesticide use (Meehan et al., 2013). Similarly, insect pollinators in urban systems may experience elevated temperatures through the urban heat effect (Hamblin et al., 2018), industrial pollution (Sivakoff et al., 2020), higher diversity and abundance of introduced species (Fitch et al., 2019; Prendergast et al., 2021) and increased pathogen pressures (Youngsteadt et al., 2015).

On the other hand, urban systems may provide opportunities for pollinator conservation (Baldock et al., 2019; Wenzel et al., 2020). The demographic rates, abundance and/or diversity of some pollinator taxa may increase along urbanization gradients (Baldock et al, 2015; Samuelson et al., 2019; Theodorou et al., 2020), although for some taxa and in other cities these general patterns do not always hold (Bates et al., 2011; Deguines et al., 2016; Theodorou et al., 2020). The variability of pollinator responses to urbanization calls for a better understanding of what makes an urban environment favourable for pollinator reproduction and survival and what types of species will benefit from different urban habitat management strategies (Baldock et al., 2019; Wenzel et al., 2020). To this end, a growing number of studies pinpoint drivers of locally observed urban pollinator biodiversity.

Undeveloped natural (or semi-natural) habitat remnants embedded in the urban landscape, such as urban greenbelts and nature reserves, may be critical for pollinators. Across many studies, locally observed pollinator abundance and diversity increases with the size of a natural habitat patch, suggesting that the presence and total area of undeveloped greenspace in an urban landscape impacts local pollinator demographics (Beninde et al., 2015). Urban bumble bee nesting densities are positively associated with the proportion of wooded habitat or forest cover in the surrounding landscape (Jha & Kremen 2013a; Conflitti et al., 2022), and losses of ground nesting wild bee species from pollinator communities are associated with high impervious surface cover (Banaszak-Cibicka & Żmihorski, 2012; Geslin et al., 2016), suggesting that pollinator nesting opportunities decline as natural habitat remnants are replaced. Given that local pollinator species composition may differ in urban natural habitat remnants (or in closer proximity to these greenspaces) versus in the developed urban matrix, these habitat patches may provide unique nesting and food opportunities that are critical for some pollinator species – potentially sensitive species with more specialized resource or habitat use requirements (Matteson et al., 2013; Martins et al., 2017; Turo et al., 2021; Prendergast et al., 2022).

Because they provide food in the form of nectar and pollen, flower resources are another key driver of local pollinator abundance and diversity (Kennedy et al., 2013; Hyjazie & Sargent, 2022). Availability of flower resources mediate the abundance and diversity of pollinators in the developed urban matrix (Simao et al., 2018; Adams et al., 2020; Cohen et al., 2021; Gerner & Sargent, 2022). At larger spatial scales, the abundance and diversity of flower resources can be difficult to directly assess, and so some studies have use proxy measures to quantify flower resources in the urban landscape including weighting the landscape based on expert opinion (Kennedy et al., 2013; Davis et al., 2017) or using remote sensing tools that capture coarse vegetation patterns (Leong & Roderick, 2015; Beduschi et al., 2018; Gonzalez et al., 2022; Casanelles‐Abella et al., 2022). Socioeconomic proxy measurements have also been used in urban biodiversity studies to represent availability of floral resources in the urban landscape (Leong et al., 2018). In both urban natural habitat remnants and in the developed urban matrix, income is a consistent predictor of plant diversity, particularly of woody-perennial plants, representing a ‘luxury effect’ (Hope et al., 2003; Martin et al., 2004; Birgimana et al., 2012; Clark et al., 2013; Kirkpatrick et al., 2013; Wang et al., 2015; Avolio et al., 2015; Nesbitt et al., 2019). Studies directly quantifying the association between income and biodiversity of mobile organisms are more limited (Leong et al., 2018). For insects, local abundance and diversity has been both positively (Leong et al., 2016; Baldock et al., 2019) and negatively (Gledhill & James 2012; Lowenstein et al., 2014) associated with neighborhood affluence, although these conclusions have been made in part without disentangling potential effects of income from potential effects of the intensity of urban development.

Despite a growing synthesis of how the urban environment shapes local patterns of pollinator abundance and diversity, we lack an understanding of whether differences among the particular configurations of urban landscapes, such as the total area of natural habitat harboured by a city or the distribution of floral resources across a city, predict patterns of biodiversity. Because dispersal among populations in a connected landscape can compensate for local population extinctions that might occur due to demographic or environmental stochasticity, ecological theory predicts that the persistence of a species as a whole in a landscape should be driven not only by the quality of local habitat experienced by individual organisms but also by the environment collectively experienced by interconnected populations (Hanski, 1999; Leibold et al., 2004). That is, that a species should be more likely to persist in a landscape as the total area and/or connectivity of favourable habitat in the landscape increases (Hanski, 1999; Leibold et al., 2004). If the landscape-scale urban environment mediates the landscape-scale persistence of a species, then isolated local habitat amendments may be insufficient to support city-wide biodiversity in the long run (Chase et al., 2020). On the other hand (describe the type of landscapes that might be supportive of higher biodiversity). By comparing landscape-scale, multi-year biodiversity metrics among urban landscapes, we can identify whether city-wide management, design and policy strategies are needed to realize long-term urban biodiversity conservation goals.

Proliferating data from natural history collections (“NHC’s”) – including both community science data and digitized museum records– presents a novel opportunity to explore whether differences among urban landscapes influence long-term, landscape-scale pollinator biodiversity patterns. Occupancy modeling approaches that account for spatial, temporal and taxonomic biases in detection ability (Mackenzie, 2006), paired with modeling choices that account for the unstructured and opportunistic nature of these data, show emerging potential to utilize NHC’s to uncover impacts of global change on broad scale biodiversity patterns (Altwegg et al., 2019; Kharouba et al., 2019; Guzman et al., 2021; Jackson et al., 2022; Shirey et al., 2022; Davis et al, 2022)

With a focus on bumble bees and hoverflies as indicators of urban pollinator biodiversity, we use NHC’s to test two hypotheses that link differences among urban landscapes to landscape-scale urban biodiversity outcomes. First, we test the hypothesis that undeveloped natural habitats embedded in the urban landscape (such as urban greenbelts and nature reserves) predict local pollinator species occurrence and, simultaneously, that greater overall area of natural habitat in the urban landscape compensates for local extinctions thereby driving higher pollinator occurrence rates at the landscape-scale. Second, using income as a proxy measure for plant diversity and management resources, we test the hypothesis that higher gridcell income favours local pollinator occurrence and, simultaneously, that higher average income across an urban landscape similarly compensates for local extinctions thereby driving higher pollinator occurrence rates at the landscape scale. We therefore predict that pollinator occupancy patterns in urban landscapes will be positively associated with total natural habitat area and average income of an urban landscape. Further, by testing for a negative correlation between species-specific effects of natural habitat area on occupancy and species-specific range wide occupancy rates, we assess the hypothesis that occurrence of rare (versus common) species is more dependent on the total natural habitat area in the urban landscape.

1. ***Methods***

***2.1 Data Collection***

***2.1.1 Urban Landscape Data***

To examine drivers of landscape-scale pollinator biodiversity, we first defined a set of urban landscapes in the continental United States. To objectively define urban landscapes, we created a 10 km by 10 km grid across the extent of the continental United States, extracted the average population density from a **300** m resolution raster layer obtained from the year 2015 (), and then filtered out any grid cells that were below a minimum population density of 1200 people per km^2. We also filtered out grid cells that ranked over 5 standard deviations above the mean population density of the remaining urban landscapes, corresponding to two sites from central New York City, which we considered to be fundamentally different from the other urban landscapes (although the main results are unchanged by this decision, Supplement \_).

Importantly, the results of the analysis are likely to be sensitive to our site selection decisions (Shirey et al., 2022), and so we regathered the data at increasing spatial scales (15 km, 20 km, and 25 km grid cell edge lengths) and decreasing minimum population density thresholds (1000, 800, and 600 people per km^2). We later conducted sensitivity analyses by refitting our models for each of these datasets and then examining the dependence of parameter estimates on the spatial scale of the study (See supplemental methods - ). We did not resolve our spatial scale more finely than 10 km by 10 km because community science contributors may optionally obscure geolocation for privacy protection (iNaturalist, ). iNaturalist coordinate obfuscation casts detections randomly within a 10 km area surrounding the contributor’s location (iNaturalist, ) and so our confidence in our ability to match detections to the correct urban landscape declines as the spatial scale of the site area declines, especially when the obfuscation area is larger than the site area. Further methods and results refer to data collection from 10 km by 10 km sites with a minimum population density threshold of 1200 people per km^2 unless explicitly stated otherwise.

Because species occurrence rates are generally expected to scale with study area (), we determined the total land area of each site within the spatial range of our NHC’s. Total land area of each site varies given that some grid cells fall on coastlines or international borders. To calculate land area, we removed 30 m x 30 m pixels from each grid cell that did not intersect with the continental U.S. international borders (Tigris reference) and/or were classified as open water (30 m resolution land cover data from 2016; NLCD reference) and then summed the number of remaining pixels. Because we consider that detection data originating finitely small sites could be a product of fewer community science contributors and thus potentially more subject to inconsistent individual contributor behaviours, we filtered out any remaining grid cells that were comprised of less than 25% land area, resulting in the final set of urban landscape sites used for each analysis (n = 459 sites) (Figure 1).

We extracted environmental covariates from each urban landscape using national land cover and socioeconomic data. We used 30 m resolution land cover data from 2016 to determine the amount of natural habitat in each urban landscape (NLCD reference), calculating the summed proportion of land area represented by any of the following undeveloped land cover classes: grassland/herbaceous cover, shrub/scrubland, forest, and wetlands. These land cover classes do not include vegetation managed for agriculture (either cropland or pasture), or open developed areas with low impervious surface cover such as low-density residential cover, managed parks and golf courses (NLCD reference), and consequently represent natural parks within the urban landscape such as nature reserves and greenbelts or other smaller herb, shrub or forest remnants (Supplemental methods \_). Next, to quantify the income for each urban landscape, we gathered census block group estimates of the median household income from the year 2020 (), where a census block groups typically represent clusters of 300 – 6000 people (). The relative purchasing power of household income (which we use as a proxy for resources that may be invested in plant management) varies across the U.S. (need ref), and so we standardized the median household income of each census block proportional to the average across all census blocks in the corresponding state. We then calculated the mean value of standardized median household incomes from all census block groups that spatially intersected each urban landscape (Supplemental methods \_).

Finally, we scaled all covariates using centered z-score standardization (Gelman et al., 20\_). Thus, the model parameter estimates described in our results represent the change in a response (detection or occupancy rate) associated with an increase in one standard deviation of a covariate.

***2.1.2 Natural History Collections Data***

Bumble bee (*Bombus*) NHC’s were obtained from the Bumble Bees of North America (BBNA) Database (Richardson, 2022). Hoverfly NHC’s were requested from GBIF (cite my download codes here) using the R package ‘rgbif’ (cite rgbif). To reduce spatial misattribution of detections, we filtered out any records lacking latitude/longitude coordinates or with coordinate uncertainty greater than 10 km. (most coordinates have a certainty of less than 1km, mean and sd = ). For the main analyses, we further filtered the detections to those collected between 2010 and 2022, constraining our comparison of pollinator biodiversity among urban landscapes within a contemporary timeframe.

Given difficult taxonomic concepts and a lack of species level identifiability (Skevington, ), we grouped all detections of *Eumerus spp.* and *Chrysogaster spp.* at the genus level and modelled these genera as single species concepts. For all other genera, including for bumble bees, we filtered out any detections lacking a resolved species-level identification. Given a recent taxonomic split (Hines et al., ) and potential lack of identification resolution in the data, we grouped and modelled bumble bee detections for *Bombus vancouverensis* and *Bombus bifarius* as a single species concept (labelled *Bombus bifarius* in the results). Because they are closely related and both contain ecologically similar, non-native species (Skevington, ), we pooled species from the sister genera *Eristalis* and *Eoseristalis* into a single genus cluster (later allowing us to model species-specific effects for this group in a shared hierarchical nesting structure).

To later model detections as arising from either of two distinct observation processes, we classified NHC’s as “community science” or “research collection” detections. For bumble bees, we classified data as community science detections if the data source provided by BBNA was labelled as any of the following: "www.bumblebeewatch.org 05-24-2017", "Bumble Bee Watch 10-26-2022", "BeeSpotter 10-03-2018", "iNaturalist 12-01-2022", "Flickr", "Xerces Society--dates various", "Xerces Society 11/14/2013", "Xerces Society 2012 Citizen Science", "Xerces Society B. franklini 02/26/2012". All other bumble bee NHC’s were classified as research collection detections. For hoverflies, we classified data as community science detections if the “basisOfRecord” was labelled as “HUMAN\_OBSERVATION” or as research collection detections if labelled as “PRESERVED\_SPECIMEN”. Incidentally, all “HUMAN\_OBSERVATION” hoverfly records from our gbif request originated from iNaturalist.

We then assigned the NHC’s to sites by intersecting their geolocation data with the urban landscape shapefile (cite sf library), filtering out any detections that fell outside of our defined urban landscapes. Because an occupancy model approach may have difficulty determining whether extremely limited detections of a species are caused by particularly narrow ecological distribution or particularly difficult detectability (), we discarded species that were detected at fewer than 3 site by year occasions (i.e., singletons and doubletons). Under the 10 km by 10 km, 1200 people per km^2 site selection criteria, we obtained \_ total detections for 32 bumble bee species concepts (\_ unique species x site x year detections) and \_ total detections for \_ hoverfly species concepts (\_ unique species x site x year detections) (Figure 2).

We used an occupancy model approach to account for spatial, temporal and taxonomic detection biases in our data that might obscure our underlying ecological quantity of interest – whether species are present in urban landscapes. To correct for detection bias, occupancy models gather information about the variability in the detection of species over temporally repeated sampling events during which the underlying ecological state of presence/absence is assumed to be fixed (Mackenzie, 2006; Kery & Royle, 2016). This requires information about non-detection events (occasions where a species was searched for but not detected) in addition to detections (occasions where a species was detected) (Mackenzie, 2006; Kery & Royle, 2016). The opportunistic detection data from the BBNA database and GBIF were not collected with the intentional design to be analyzed with an occupancy model and consequently these data required further preliminary processing. Specifically, we needed to explicitly define the duration of temporally replicated sampling events within periods of ecological closure and also infer non-detections (Shirey et al., 2022).

Here, we defined our sampling events as individual years contained within closed occupancy intervals (as in Jackson et al., 2022; Shirey et al., 2022), dividing the full twelve-year time period into 4 occupancy intervals (each consisting of 3 temporally replicated sampling events) as our baseline (Figure 2). Following Shirey et al., 2022, we later conducted sensitivity analyses by refitting our models under alternative temporally defined occupancy intervals (6 sets of 2 sampling events and as 3 sets of 4 sampling events) and examining the change in parameter estimates (See supplemental methods - ).

We inferred non-detections in our data by implying community surveys (as in Jackson et al., 2022; Shirey et al., 2022). To imply community surveys for bumble bees we grouped NHC data by year, site and the name of the observer/s provided by BBNA. If two or more bumble bee species were detected by the same observer/s at a site in a year, then we implied that a community survey event occurred. For hoverflies, we grouped data by year, site and institution (representing a unique digitized collection that maintains the specimen). If two or more hoverfly species were detected by the same institution at a site in a year, we then implied that a community survey event occurred. For each taxonomic group, when a community sampling event was inferred, all undetected species were then scored as non-detections. If no community sampling event was inferred for a site/year, then we later prevented data from contributing to the model likelihood function (i.e., data treated similar to NA values; see Section 2.2 Statistical Analyses). Single species detections were retained without inferring non-detections of other species in the community. In short, we treat research collections community sampling events as a binary observation phenomenon – the community at a site is sampled in some years but not in others.

For community science data, we instead imply a continuous full-community sampling process. That is, at these spatial scales and in these highly active urban areas (cf. proportion of all records that come from urban sites versus proportion of land cover represented by urban sites), we assume that all species are searched for and could possibly be detected in any given year. We assume that any variation in detection rates can be explained by site covariates such as year (potentially increasing through time as overall iNaturalist observations increase) or population (potentially increasing in areas with more people who might act as contributors) and by species-specific detection properties (some species consistently more difficult to detect versus others). At 10 km by 10 km scales, many observers contribute to the observation process (mean number of observers per site for our urban landscapes in 2022 = \_; sd = \_). . This volume of contributors reduces the potential for site detection rates to be driven by the behaviours of any single contributor. For example, one contributor may randomly disregard contributing observations of a species that they find uninteresting. But with \_1000 other contributors scouring the site, another individual might find a particular interest in that species, and so on average their peculiarities might cancel out. If instead a species is rarely detected at a site because all contributors generally show low interest in it or occurs at low densities that make it difficult to observe, and that pattern holds across many sites, then this will be captured in the species-specific effect that can be included in our modelling approach. Identification of specimens, the second hurdle required to move an observation from iNaturalist to GBIF, is also unlikely to be effected by site specific peculiarities. Most iNaturalist observations are identified by a small subset of users that inordinately contribute to observations from across the country, not necessarily tied to a certain site (mean versus median detections?), meaning that identification is less driven by site specific heterogeneity and more so by species effects (i.e., is the species difficult for the core community of contributors to identify from photographs).

Finally, before analyzing our data, we inferred whether species could realistically occur at each of the sites (Shirey et al., 2022). We mapped the distribution of each species by drawing a convex hull around all our NHC detections from 2000 to present (as in Jackson et al., 2022; Shirey et al., 2022). We intersected the urban landscape shapefile with each of these species distributions (Supplemental Figure) (Jackson et al., 2022; Shirey et al., 2022). We then prevented data from species/site combinations outside of each range from contributing to the model likelihood function (i.e., data treated similar to an NA value; see Section 2.2 Statistical Analyses). This approach likely overestimates the range of some species and underestimates the range of others, but on average provides a quantitative metric of whether a species could reasonably occur in each urban landscape (). We manually filtered the ranges for two species: For *Bombus impatiens*, we filtered out detections west of -100 degree longitude where it is not native but has been recently introduced via commercial greenhouse use (ref); For the federally endangered and putatively declining *Bombus affinis*, we filtered detections to the upper Midwest states – it was detected once in New Mexico, and several times in New England between 2000-2005 but has not been recorded in our data set in these regions since. See supplementary figures. We further filtered our data set removing any species that did not occur 5 or more times since 2000 so that we only include species for which we can confidently draw a distribution map.

***2.2 Statistical Analyses***

***2.2.1 Integrated Occupancy Models***

We compared pollinator occupancy rates among urban landscapes using a multi-species, integrated hierarchical Bayesian modeling framework, which extends the classic occupancy model approach (Dorazio, 2014 check ref), allowing the different observation processes of community science and traditional research-based collections to concurrently contribute to our understanding of the shared underlying ecological state of interest (species occurrence). An integrated model form allows greater flexibility when including NHC data from these different sources, which may be governed by different processes, and by including all information available an integrated model framework can estimate ecological parameters more precisely than if estimated from each observation process alone (Dorazio, 2014; Zipkin et al., 2021; Davis et al., 2022). Multi-species occupancy modelling partially-pools data among species, allowing the variation in detection and occurrence rates from the community to inform estimates for species-specific trends ().

Here, we model detections (the number of years in an interval where a species was detected at least once by a NHC observation process) of a species at site during occupancy interval as conditional on occupancy state , where occupancy state is the outcome of a Bernoulli distributed random variable:

Detections are then modelled as the stochastic outcome of a binomial trial of size , where indicates the number of years in the occupancy interval. Where community science () and research collections () are conditional on the shared underlying latent (only partially observed) variable :

1. ; and

We account for the binary process of sampling in our research collections NHC’s by reducing for any years in which a community sampling event (or species-specific sampling event for species ) was not inferred in the data preparation stage. Within a Bayesian HMC simulation implemented using STAN (), we estimated the likelihood of the data given for any site/species/interval combinations where there is at least one detection in either NHC dataset, or the marginal likelihood of the data given (species occurs at site/interval but was detected) and given (species does not occur at the site/interval) for any site/species/interval combinations with no detections.

We modeled heterogeneity in occurrence rates using a logit-link for the Bernoulli rate , including a global occurrence intercept (); phylogenetically-nested, species-specific intercept effects () and spatially-nested, site-specific intercept effects (); and predictor variables including site land cover area (), species-specific effect of natural habitat area (), and species specific effect of average income ():

Where, species-specific intercepts are centered on a global intercept, for bumble bees (all from the same genus):

And, for hoverflies (from \_ different genera):

For both taxonomic groups, site-specific intercept effects are not centered on the global intercept and are nested as:

And for both taxonomic groups, where indicates the community mean and indicates scale, species-specific effects on predictors are specified as:



We modeled heterogeneity in detection rates also using a logit-link for the binomial success rate , including a global occurrence intercept (); phylogenetically-nested, species-specific intercept effects () and spatially-nested, site-specific intercept effects (). For community science NHC’s we included detection predictors corresponding to our hypotheses that detection rates would increase in areas with more people (because this equates to more potential contributors) and exponentially through time (because community science contribution rates have exponentially increased through time (Figure 2)) – and , respectively; For research collection NHC’s we included a detection predictor corresponding to a hypothesis that detection rates are higher when there are more detections (a proxy for sampling effort) – , where total records are averaged across all years with community sampling events in an occupancy interval:

Similarly to our model for the ecological process, species-specific intercept effects were phylogenetically nested and site-specific intercept effects were spatially nested, using the same hierarchical nesting structure for both and . Where, species-specific intercepts are centered on a global intercept, for bumble bees (all from the same genus):



And, for hoverflies (from \_ different genera):

For both taxonomic groups, site-specific intercept effects are not centered on the global intercept and are nested as:



Subregions were determined by intersecting the urban landscape sites with level 3 U.S. Ecoregions () and regions by intersecting the sites with level 1 U.S. Ecoregions () (Supplemental Figure \_). The use of this multi-level clustering structure simultaneously allows us to account for the temporal pseudoreplication in our data and also accounts for unmodeled heterogeneity that is likely to be spatially autocorrelated – i.e., that there may be environmental or management similarities relevant to pollinators in urban landscapes clustered within the same metropolitan area, and that there may be environmental or management similarities relevant to pollinators in metropolitan areas clustered within the same broad region.

To estimate our main parameters (i.e., the community mean effect of natural habitat on occurrence, the community mean effect of income on occurrence), we used weakly-informative priors to inform the posterior likelihood of the data given the parameter values generated during each step of the HMC implementation. Specifically, we use priors for these parameters, which after being transformed back from a logit scale to a probability scale equate to 95% prior probability distribution of the effect of change in 1 standard deviation of the covariate is centered on 50% (does not change occurrence rate to deviate from random chance) and ranges from reducing to 1.8% or increasing to 98.2% (assuming all other effects being 0). Thus this type of weak prior centers on zero effect of the covariate and limits effects from approaching -/+ . To enable model identifiability given sparse data, particularly for the museum observation process,

To assess goodness of fit of the model outputs we used Tukey-Freeman p-value (TFP) posterior predictive checks, binning data by species and comparing the detections in the real data against detections for data sets stochastically generated using our parameter estimates during each iteration of the HMC implementation. The summary FTP indicates the proportion of times that the real data versus the generated data are closer to the expected values, with FTP < 0.1 indicating a serial lack of model fit.

We used a data simulation approach to confirm the validity of our model under the data assumptions outlined here. Specifically, we 1) used known parameter values to simulate stochastic pollinator communities in urban landscapes (using similar sample sizes and similar proportions of missing data), 2) fit each model to the simulated data, and 3) confirmed that 95% Bayesian Confidence Interval from the posterior distribution generated by the HMC implementation overlaps with the known values used to generate the data (Supplemental Figure \_).

***2.2.2 Species specialization correlations***

**Software:**

Analyses were conducted using STAN version 2.21.0 (ref), implemented through the R interfacing package “rstan” using R version (). Data and code are available at: https://github.com/jensculrich/occupancy\_model\_for\_urban\_NHC\_records

***2.2.1 Urban landscape occupancy model***

***2.2.2 Correlations with habitat specialization***

***2.2.3 Sensitivity analyses.***

***3. Results***

Save for modelling decisions:

For Syprhidae I requested data for all available preserved specimens and human observations identified to the species level and which contained latitude and longitude coordinates, including specimens from any year. I ensured that the uncertainty in the coordinates was less than the spatial extent of the site area, by….. These initial filters result in a dataset comprising 34,217 unique observation records. All human observations are research-quality observations contributed from iNaturalist (20153 records). Preserved specimens (14,064 records) originate from multiple sources, most prominently including the Natural History Museum of Los Angeles County (LACM) (10,413 records) and the Randall Morgan Insect Collection at the University of California, Santa Cruz (UCSC) (2,347 records). 85.3% of these 34,217 records originate from the year 2000 or later, representing 137 species, including the most common *Toxomerus marginatus* (>3,500 records since year 2000 or later, Figure X).

Chart

Description automatically generated

***Figure X: Number of records for hover fly species with 100 or more records collected in California since the year 2000.***

Separate occurrence arrays for iNaturalist data and museum record data are then produced (where presence/absence of species at sites in an interval at a visit is denoted). These separate detection data arrays are handled differently by the model using an integrated hierarchical approach (Zipkin et al., 2021; Davis et al., 2022)

***4. Discussion***

***4. Discussion***

**First Discussion paragraph:** While previous studies have identified drivers of locally observed urban pollinator biodiversity, to our knowledge this is the first study to use an across city comparison to test whether two proxies for pollinator habitat (natural area and neighbourhood income) predict biodiversity patterns in urban areas across broad spatial and temporal scales. Individual bumble bee species showed wide variation in their responses to natural habitat area in the urban landscape, however, our results also indicate that the occupancy rate of the average bumble bee species does not have a strong average association with total natural habitat area. On the other hand, we identified a positive association between bumble bee occupancy rates and the relative income of a grid cell (a proxy measure for plant diversity and vegetation management resources [cite evidence here, give foreshadowing in intro]), highlighting that investment in management of the vegetation in the matrix of the anthropogenic landscape itself, separate from inclusion of large areas of undeveloped habitat, is a key component of urban pollinator conservation. Interestingly, we found the opposite pattern for hoverflies, with hoverflies exhibiting a positive association between total natural habitat area and landscape-scale occupancy, but no impact of income (right?). This result emphasizes that natural habitat remnants such as urban greenbelts and nature reserves remain essential for the persistence of certain pollinator taxa. Together, our results demonstrate that drivers of local pollinator abundance and diversity also operate at larger spatial and temporal scales, mediating landscape-scale urban pollinator occupancy – although with different drivers and varying effect sizes among species and taxonomic groups. We discuss the assumptions and limitations of our results as well as their implications for urban biodiversity conservation applications and, moreover, for our understanding of the spatiotemporal scale dependence of biodiversity patterns.

*Some discussion ideas:*

assumptions and caveats of the occurrence data and ecological quantities. Occurrence to species richness and occurrence to abundance

assumptions of income as a proxy and socioeconomic implications. Variability in ‘natural habitat’

discussion of detection parameter estimates and role of community science data

relationship between local and landscape biodiversity dynamics. Effects diminish as spatial scale increases

applications – city wide management, can’t add habitat but can limit further reductions? What is/could be done to protect urban remnant habitats. What is/could be done to redistribute resources among urban landscapes from areas with wealth to those without?

Consequently, management plans and decisions that aim to benefit pollinators and maintain ecosystem services will compound benefits from actions and policies that take a city-wide scope.

**Last Discussion paragraph:** Much of urban ecology and conservation science, and ecology generally, is focused on quantifying biodiversity metrics at a local scale, assuming that local populations and communities are spatially closed and isolated (Leibold et al., 2004; Chase et al., 2020). With this study we examined relationships between the environment and species occupancy at multiple broader spatial scales to test the hypothesis that city-wide landscape quality mediates city-wide pollinator biodiversity. For hoverflies, urban landscapes with larger amounts of natural habitat area sustain groups of interconnected populations, while for bumble bees, affluent urban landscapes with presumably more resources to invest in vegetation cover and management – including in the developed urban matrix itself – sustain groups of interconnected populations. Thus, the fate of a species in the landscape is mediated not just by the local demography of a subpopulation, but also by the broader environment collectively experienced by all subpopulations (as is suggested by metapopulation and metacommunity perspectives) (Leibold et al., 2004; Chase et al., 2020). The properties of the broader overall landscape may influence landscape-scale species persistence by moderating the degree to which subpopulations can compensate for local extinctions due to environmental or demographic stochasticity (Leibold et al., 2004; Chase et al., 2020). The conclusions from this cross-landscape analysis call for increased consideration of the interplay between local population and community dynamics with landscape to regional dynamics. In application, this demands that local urban habitat restoration and enhancements coordinate with city-wide policy and planning to ensure long-term success of species conservation.

***Other notes***

, potentially focusing efforts to support plant diversity management in residential land use areas given that these comprise the largest spatial proportion of north American cities. Efforts to densify existing developed areas in to allow urban population to grow without removing natural habitat remnants?

not only by the local quality experienced These cross-landscape analyses that find that occupancy of an entire metapopulation hinges on the environment experienced by the entire metapopulation support the idea that local populations and communities are not closed and isolated but rather are interconnected to other populations and communities in the shared landscape. but a larger amount of natural

more area of natural habitat reduces the odds that an entire metapopulation will be lost from the urban environment.

function as connected metapopulation units. In the case of hoverflies, that more area of favourable habitats in an urban landscape reduces overall extinction a shared landscape

– making progress towards understanding how the local interacts with the landscape.

***References***

Adams, B. J., Li, E., Bahlai, C. A., Meineke, E. K., McGlynn, T. P., & Brown, B. V. (2020). Local‐and landscape‐scale variables shape insect diversity in an urban biodiversity hot spot. *Ecological Applications*, *30*(4), e02089.

Aizen, M. A., Smith‐Ramírez, C., Morales, C. L., Vieli, L., Sáez, A., Barahona‐Segovia, R. M., ... & Harder, L. D. (2019). Coordinated species importation policies are needed to reduce serious invasions globally: The case of alien bumblebees in South America. *Journal of Applied Ecology*, *56*(1), 100-106.

Aronson, M. F. J. et al 2014. A global analysis of the impacts of urbanization on bird and plant diversity reveals key anthropogenic drivers. Proceedings of the Royal Society B 281:20133330.

Beduschi, T., Kormann, U.G., Tscharntke, T., and Scherber (2018). Spatial community turnover of pollinators is relaxed by semi-natural habitats, but not by mass-flowering crops in agricultural landscapes. *Biol. Conserv.* 221, 59–66. doi: 10.1016/j.biocon.2018.01.016

Brittain, C. A., Vighi, M., Bommarco, R., Settele, J., & Potts, S. G. (2010). Impacts of a pesticide on pollinator species richness at different spatial scales. *Basic and Applied Ecology*, *11*(2), 106-115.

Casanelles‐Abella, J., Müller, S., Keller, A., Aleixo, C., Alós Orti, M., Chiron, F., ... & Moretti, M. (2022). How wild bees find a way in European cities: Pollen metabarcoding unravels multiple feeding strategies and their effects on distribution patterns in four wild bee species. *Journal of Applied Ecology*, *59*(2), 457-470.

Colla SR, Otterstatter MC, Gegear RJ, Thomson JD (2006) Plight of the bumble bee: pathogen spillover from commercial to wild populations. Biol Conserv 129:461–467

Conflitti, I. M., Arshad Imrit, M., Morrison, B., Sharma, S., Colla, S. R., & Zayed, A. (2022). Bees in the six: Determinants of bumblebee habitat quality in urban landscapes. *Ecology and Evolution*, *12*(3), e8667.

Cheptou, P. O., & Avendaño V, L. G. (2006). Pollination processes and the Allee effect in highly fragmented populations: consequences for the mating system in urban environments. *New Phytologist*, *172*(4), 774-783.

Davis, A. Y., Lonsdorf, E. V., Shierk, C. R., Matteson, K. C., Taylor, J. R., Lovell, S. T., & Minor, E. S. (2017). Enhancing pollination supply in an urban ecosystem through landscape modifications. *Landscape and Urban Planning*, *162*, 157-166.

Fürst, M. A., McMahon, D. P., Osborne, J. L., Paxton, R. J., & Brown, M. J. F. (2014). Disease associations between honeybees and bumblebees as a threat to wild pollinators. *Nature*, *506*(7488), 364-366.

Gledhill DG, James P. 2012 Socio-economic variables as indicators of pond conservation value in an urban landscape. Urban Ecosyst. 15, 849–861. (doi:10. 1007/s11252-012-0242-7)

Gonzales, D., Hempel de Ibarra, N., & Anderson, K. (2022). Remote Sensing of Floral Resources for Pollinators–New Horizons From Satellites to Drones. *Frontiers in Ecology and Evolution*, *10*.

Grimm, N. B., S. H. Faeth, N. E. Golubiewski, C. L. Redman, J. Wu, X. Bai, and J. M. Briggs. 2008. Global change and the ecology of cities. Science 319:756–760.

Hamblin, A. L., Youngsteadt, E., & Frank, S. D. (2018). Wild bee abundance declines with urban warming, regardless of floral density. *Urban Ecosystems*, *21*, 419-428.

Hope D, Gries C, Zhu WX, Fagan WF, Redman CL, Grimm NB, Nelson AL, Martin C, Kinzig A. 2003 Socioeconomics drive urban plant diversity. Proc. Natl Acad. Sci. USA 100, 8788–8792. (doi:10.1073/ pnas.1537557100)

Jha, S., Egerer, M., Bichier, P., Cohen, H., Liere, H., Lin, B., ... & Philpott, S. M. (2023). Multiple ecosystem service synergies and landscape mediation of biodiversity within urban agroecosystems. *Ecology Letters*, *26*(3), 369-383.

Jha, S., & Kremen, C. (2013a). Resource diversity and landscape-level homogeneity drive native bee foraging. *Proceedings of the National Academy of Sciences of the United States of America*, **110**, 555–558. <https://doi.org/10.1073/pnas.1208682110>

Jha, S., & Kremen, C. (2013b). Urban land use limits regional bumble bee gene flow. *Molecular Ecology*, **22**, 2483–2495 <https://doi.org/10.1111/mec.12275>.

Kaiser-Bunbury, C. N., Mougal, J., Whittington, A. E., Valentin, T., Gabriel, R., Olesen, J. M., & Blüthgen, N. (2017). Ecosystem restoration strengthens pollination network resilience and function. *Nature*, *542*(7640), 223-227.

Kerr, J.T. & Cihlar, J. (2003). Land use and cover with intensity of agriculture for Canada from satellite and census data. *Global Ecol. Biogeogr.*, **12**, 161 – 172.

Klein Goldewijk, K. 2001 Estimating global land use change over the past 300 years: the HYDE database. Global Biogeochem. Cycles 15, 417–433.

Leitão, R. P., J. Zuanon, S. Villéger, S. E. Williams, C. Baraloto, C. Fortunel, and D. Mouillot. 2016. “Rare Species Contribute Disproportionately to the Functional Structure of Species Assemblages.” *Proceedings of the Royal Society B: Biological Sciences* **283**(1828): 20160084.

Leong, M., and Roderick, G. K. (2015). Remote sensing captures varying temporal patterns of vegetation between human-altered and natural landscapes. *Peerj.* 3:e1141. doi: 10.7717/peerj.1141

Leong M, Bertone MA, Bayless KM, Dunn RR, Trautwein MD. 2016 Exoskeletons and economics: indoor arthropod diversity increases in affluent neighbourhoods. Biol. Lett. 12, 20160322. (doi:10. 1098/rsbl.2016.0322)

Lin, B. B., Philpott, S. M., & Jha, S. (2015). The future of urban agriculture and biodiversity-ecosystem services: Challenges and next steps. *Basic and applied ecology*, *16*(3), 189-201.

Martins, K. T., Gonzalez, A., & Lechowicz, M. J. (2017). Patterns of pollinator turnover and increasing diversity associated with urban habitats. *Urban Ecosystems*, *20*, 1359-1371.

Matteson, K. C., Grace, J. B., & Minor, E. S. (2013). Direct and indirect effects of land use on floral resources and flower‐visiting insects across an urban landscape. *Oikos*, *122*(5), 682-694.

McDonald, R. I., Biswas, T., Sachar, C., Housman, I., Boucher, T. M., Balk, D., ... & Leyk, S. (2021). The tree cover and temperature disparity in US urbanized areas: Quantifying the association with income across 5,723 communities. *PloS one*, *16*(4), e0249715.

McKinney, M. L. 2006. “Urbanization as a Major Cause of Biotic Homogenization.” *Biological Conservation* **127**(3): 247–60.

Meehan, T. D., Werling, B. P., Landis, D. A., & Gratton, C. (2011). Agricultural landscape simplification and insecticide use in the Midwestern United States. *Proceedings of the National Academy of Sciences*, *108*(28), 11500-11505.

Moroń, D., Lenda, M., Skórka, P., Szentgyörgyi, H., Settele, J. & Woyciechowski, M. (2009) Wild pollinator communities are negatively affected by invasion of alien goldenrods in grassland landscapes. *Biological Conservation*, **142**, 1322–1332.

Nesbitt, L., Meitner, M. J., Girling, C., Sheppard, S. R., & Lu, Y. (2019). Who has access to urban vegetation? A spatial analysis of distributional green equity in 10 US cities. *Landscape and Urban Planning*, *181*, 51-79.

Oliver, T. H., Marshall, H. H., Morecroft, M. D., Brereton, T., Prudhomme, C. & Huntingford, C. (2015). Interacting effects of climate change and habitat fragmentation on drought-sensitive butterflies. *Nature Climate Change* **5**(10), 941–945.

**Richardson, L. L. 2022. Bumble bees of North America occurrence records database.**[***https://www.leifrichardson.org/bbna.html*. Data accessed 02-15-2023**](https://www.leifrichardson.org/bbna.html.%20Data%20accessed%2002-15-2023)***.***

Roth, M., Oke, T. R., & Emery, W. J. (1989). Satellite-derived urban heat islands from three coastal cities and the utilization of such data in urban climatology. *International Journal of Remote Sensing*, **10**(11), 1699–1720. [https://doi.org/10.1080/014311689089040](https://doi.org/10.1080/01431168908904002)

Seto, K. C., B. G¨uneralp, and L. R. Hutyra. 2012. Global forecasts of urban expansion to 2030 and direct impacts on biodiversity and carbon pools. Proceedings of the National Academy of Sciences USA 149:16083–16088.

Simao, M. C. M., Matthijs, J., & Perfecto, I. (2018). Experimental small‐scale flower patches increase species density but not abundance of small urban bees. *Journal of applied ecology*, *55*(4), 1759-1768.

Simkin, R. D., Seto, K. C., McDonald, R. I., & Jetz, W. (2022). Biodiversity impacts and conservation implications of urban land expansion projected to 2050. *Proceedings of the National Academy of Sciences*, *119*(12), e2117297119.

Theodorou, P., Radzevičiūtė, R., Lentendu, G., Kahnt, B., Husemann, M., Bleidorn, C., ... & Paxton, R. J. (2020). Urban areas as hotspots for bees and pollination but not a panacea for all insects. *Nature communications*, *11*(1), 576.

Tscharntke, T., Tylianakis, J. M., Rand, T. A., Didham, R. K., Fahrig, L., Batáry, P., … Westphal, C. (2012). Landscape moderation of biodiversity patterns and processes - Eight hypotheses. Biological Reviews, 87(3), 661–685.

United Nations, Department of Economic and Social Affairs, Population Division. 2018. “World Urbanization Prospects: The 2018 Revision.”

Valtonen, A., Jantunen, J. & Saarinen, K. (2006) Flora and lepidoptera fauna adversely affected by invasive *Lupinus polyphyllus* along road verges. *Biological Conservation*, **133**, 389–396.

Verboven, H. A., Uyttenbroeck, R., Brys, R., & Hermy, M. (2014). Different responses of bees and hoverflies to land use in an urban–rural gradient show the importance of the nature of the rural land use. *Landscape and Urban Planning*, *126*, 31-41.

**2b) Statistical Analysis:**

The effect of differences between differences between cities on pollinators is approached using an integrated hierarchical modeling framework for natural history collections (NHCs). NHC data were not collected to address questions about population trends at large scales, and therefore several modeling considerations must be included to account for the opportunistic, presence only nature of these data (Shirey et al. 2022, Jackson et al. 2022). An integrated model form allows greater flexibility when including NHC data from different sources, which may emerge from very different underlying collection processes (Davis et al. 2022). By including all information available, an integrated model framework can estimate population parameters more precisely than if estimated from each observation process alone (Zipkin et al. 2021, Davis et al. 2022). “Integrated modelling also provides a robust framework by which to capitalize on the individual strengths of different data types while minimizing their weaknesses. For example, presence- only data are easier to collect across broad spatial and temporal extents than standardized data but information- poor relative to structured counts. Integrating presence- only and structured survey data can expand the spatiotemporal scope of inference while allowing for the robust estimation of parameters that cannot be estimated using presence- only data (Dorazio, 2014; Farr et al., 2020).” Quote from Davis et al. 2022.

**I am currently developing the model by slowly increasing the complexity, and will update here as revisions and improvements are made.**

*model\_intergated.stan:* Models the effects of between city differences on pollinator occupancy to differences between cities. Here, we assume that if at least one species was collected from the site\* interval\*visit, then a community survey has occurred during the site\* interval\*visit. Here we also assume that all species may occur at any site included in the study. An updated final version of the model should especially be updated to only model pollinator occupancy within the range of the pollinator species. I.e., we should not expect to see a species restricted to southern California in northern California, and if we do not restrict the analysis for this species to southern California, then we may underestimate detection probability.

Diagram

Description automatically generated Chart

Description automatically generated

**Figure X:** Analysis is constrained to sites within the range of each species. Species ranges were inferred by drawing a convex hull around all occurrence records from 1970-present. This method likely underestimates the range of some species and overestimates the range of others but on average provides a fairly realistic representation of sites where a species could in theory exist given its physiological requirements, interaction partners, general nesting substrate requirements, and long-distance dispersal limitation.

**High level model structure**

The detection data is the product of an ecological process, and an observation process that is conditional on the outcome of the ecological process. Here, two different natural history collection (NHC) datasets with their own unique observational processes give us information about the shared latent state variable on which they are both conditional. We integrate information from both NHC sources to provide better estimates of the ecological process while allowing the data to emerge uniquely from these different sources.

Ecological process:

Integrated observation process:

1. ; and

*Where:*

refers to a species.

refers to a site.

refers to an occupancy interval (set of years where occupancy state is assumed to be held constant).

refers to a visit (a sampling year within an occupancy interval. To estimate values of , the total number of visits within each occupancy interval must be >= 2).

is the latent occupancy state.

is the probability that a site is occupied.

is the number of successful detections by citizen science across all sampling rounds (*l*) within an interval (*k*)at a site (*j*)for a species (*i*).

is the number of successful detections by sampling deposited in museum records.

is the maximum number of times that a successful detection may occur within an interval. may be less than the number of observation sampling rounds within an interval if museum record metadata indicates that no community surveys were conducted during a round. Community sampling is inferred to have occurred for all sites in all time periods for the citizen science data set. is contracted to 0 for all time intervals if the site is not within the species range. is assumed to be equivalent to the total number of sampling rounds within an interval if the site is in the range and community sampling has occurred.

is the probability of detection within a sampling visit via the citizen science sampling process.

is the probability of detection within a sampling visit via the museum record sampling process.

**Low level model structure:**

*Ecological process*

*Where:*

is a global, grand-mean occupancy intercept.

, and represents a species-specific intercept.

, where is the mean effect of population density across species and describes the variation across species.

is the (scaled) average human population density within the urban site *j*. (see data details for how this information was collected and how it was transposed onto the site given the site grid parameters specified in the run\_model.R program).

is a fixed effect of spatial extent of the site. Some sites are not completely overlapping with a) land and b) the administrative area from which NHC data were drawn. A neutral process would suggest lower likelihood of occurrence in smaller sites.

is the (scaled) spatial area of the site *j* in m². (see data details for how this information was collected and how it was transposed onto the site given the site grid parameters specified in the run\_model.R program).

, where is the mean effect of population density across species and describes the variation across species.

, where is the mean effect of population density across species and describes the variation across species; and

is the temporal interval during which an observation was/was not made.

*Observation process (citizen science data)*

*Where:*

is a global, grand-mean occupancy intercept.

, and represents a species-specific intercept.

, and represents a site-specific intercept.

is a fixed effect of population density on detection rate (i.e., more people means potentially more iNaturalist contributors).

is the (scaled) average human population density within the urban site *j*. (see data details for how this information was collected and how it was transposed onto the site given the site grid parameters specified in the run\_model.R program).

is a fixed effect of time on detection rate; and

is the temporal interval during which an observation was/was not made.

*Observation process (museum record data)*

is a global, grand-mean occupancy intercept.

, and represents a species-specific intercept.

, and represents a site-specific intercept.

is a fixed effect of population density on detection rate (i.e., more people means potentially more iNaturalist contributors).

is the (scaled) average human population density within the urban site *j*. (see data details for how this information was collected and how it was transposed onto the site given the site grid parameters specified in the run\_model.R program).

is a fixed effect of time on detection rate; and

is the temporal interval during which an observation was/was not made.

**Software:**

Analyses were conducted using STAN version 2.21.0 (ref), implemented through the R interfacing package “rstan” using R version (). Data and code are available at: https://github.com/jensculrich/occupancy\_model\_for\_urban\_NHC\_records

**Prior Predictive Checks:**

The prior distributions were chosen to be wide and weakly informative, intended to provide biologically reasonable starting values: (cauchy(0, 2.5) for all parameter means and half-cauchy(0, 2.5) for all variance parameters, which is a general standard for weak priors for logistic regression (). The priors will be widened by a factor of 3 and the results rexamined to confirm that the priors have not strongly influenced the results.

**Simulation Checks:**

Prior to application on real data, the model is fit to data that is simulated from known parameter distributions to confirm that the model has a reasonable ability to capture true parameter values within the output estimates.

**Model diagnostics:**

The model outputs are assessed to confirm that there is sufficient mixing of the chains with no/low divergent transitions. The model is set to run for n=1,000 iterations with ½ n discarded as burnin, and an adapt delta value of 0.8. These values can all be upped for the real model run, but just leaving short for now while I work out the kinks.

**Posterior Predictive Checks:**

A posterior predictive check will be conducted to validate the goodness of fit of the model. The Post-PC will be added as a generated quantity automatically output by the model, and will generally follow the form presented here: https://mc-stan.org/users/documentation/case-studies/dorazio-royle-occupancy.html.

**Sensitivity Analyses:**

* Sensitivity analysis should be conducted for the spatial grain. The default value is set at 25km X 25km. I will explore if and how the results change at 10km X 10km and 50km X 50km. Ideally, the large amount of iNaturalist data will help us narrow to a smaller spatial grain so that we can more closely tie individual pollinator occurrences with landscape features in a closer radius/more pertinent biological scale. The drawback of going to a finer scale is that a) we have less data per site informing our detection rates and b) peculiarities in individual collector behaviours may cause more observation heterogeneity between sites that are ecologically equivalent.
* Sensitivity analysis should be conducted for the occupancy interval grain. The default value is set at 3 years. I will examine how shorter (2 years) or longer (up to say 5-7 years) division of intervals impacts the results.
* Sensitivity analysis should be conducted on the era start/end dates. We will make use of data up to 2022. The data were pulled from the year 2000 and later. iNaturalist data does not begin until the year 2008. Any analyses starting before 2008 should split the observation process.
* Sensitivity analysis should be conducted on the minimum population density for a site to be included and the minimum number of records per species for the species to be included.

1. **Preliminary Results**
2. **Ideas for expansion/different approaches:**

* Model separately, bumble bees and hover flies. Do the two taxonomic groups respond the same? Can we include a species trait and ask how this mediates the ecological process? Do we expect the two groups to respond similarly or differently and why?
* Another output could be a table including the mean and 95% interval for occupancy rates and affects for all taxa. Which species are most common/most rare? Which species respond most positively/negatively to management? Which species are changing most positively/negatively with respect to time?
* Increase the scale beyond California, say all west coast states (include NV, OR, WA), southwestern states (include NV, AZ, NM, UT, CO), western states (all of the above and broader).
* Reframe the question to focus on relative abundance, with citizen science data representing encounters mediated via a thinning process while museum records still being a binary presence/absence outcome BUT where presence is higher where abundance is higher. See figure below for monarch butterflies case study (Davis et al. 2022):
* Graphical user interface, text, application

  Description automatically generated

*Using natural history records to determine the effects of city-wide management differences on long-term, landscape-scale pollinator population dynamics.*

Jens Ulrich, started Nov 22, 2022.

**Contents:**

1. **Background / motivation**
   1. **Context**
   2. **Goals**
   3. **Hypotheses and predictions**
2. **Methods**
   1. **Data collection**
      1. *Data Preparation Overview*
      2. *NHC records*
      3. *Spatial data*
   2. **Statistical Analysis**
      1. Higher-level model structure
      2. Lower-level model structure
      3. Software
      4. Prior predictive checks
      5. Simulation
      6. Model diagnostics
      7. Posterior predictive checks
      8. Sensitivity analyses
3. **Preliminary Results** haven’t included anything here yet in this doc
4. **Ideas for expansion/different approaches**
5. **Background / motivation**

**1a) Context**

Currently, \_ % of the global population lives in urban areas (), and urban land use will continue to increase in the coming decades, replacing natural habitat at broad landscape scales (Guneralp & Seto, 2013; U.N., 2018). As this global land use change process continues, it is critical that conservation actions and land development plans that maximize the value of urban areas for biodiversity are implemented, thereby minimizing population decline and species loss as natural habitats are replaced.

Pollinators provide a model system for examining how to manage urban landscapes for biodiversity. Studies of pollinator populations and communities along urbanization gradients show conflicting evidence for the effects of urbanization as a whole, indicating that not all cities and towns are the same for all pollinator taxa – some urban landscapes are pollinator hotspots whereas others act as agents of declines, and some pollinator species generally respond well to urban land use change whereas others fare poorly (Wenzel et al., 2020; Fenoglio et al., 2021). Quantitative comparisons of pollinator populations and communities among urban habitats may be able to shed light on what makes a city good for biodiversity and, further, what types of species may require particular

Many small-scale studies look within a city to examine how local pollinator abundance and diversity are affected by differences in the local environmental conditions such as temperature (Adams et al., 2020) or flower resources (Adams et al., 2020; Gerner and Sargent 2021), and by the landscape context including the amount of impervious surface cover and tree cover within the foraging area of local pollinator collection (Adams et al., 2020; Gerner and Sargent 2021, Hyjazie and Sargent 2022). From these studies we know that local pollinator abundance and diversity is especially tied to the local flower abundance and species richness and local proportion of native flowering plants (Adams et al., 2020; Gerner and Sargent 2021). Differences in the local pollinator abundance and diversity can also be explained by the local habitat type, which captures variation in multiple local and landscape variables (Normandin et al. 2017, Baldock et al. 2019). Meta-analysis also tells us that, across studies that collect pollinators from semi-natural areas embedded in an urban matrix, ‘patch size’ of the semi-natural area where pollinators are collected (defined as?) has the largest effect on the local abundance and diversity of those pollinators (relative to other common explanatory variables such as… ?) (Beninde et al. 2015) .

Although these types of studies clearly show that local habitat quality can influence the amount and types of pollinators locally observed, it remains unclear whether these small local habitat qualities alter the population dynamics of pollinators at larger scales. Changes to local habitat, like the addition of flower enhancements, could simply redistribute pollinators on the landscape rather than increase survival and reproduction (Kleijn et al. 2006). Alternatively, local changes to survival and reproduction due to local habitat quality could be trivial to population and community dynamics if long-term dynamics are driven by a combination of local dynamics with dispersal between populations and communities (Hanski 1998; Leibold et al., 2004), as is observed at least in spatially-explicit agricultural pollinator systems (Ponisio et al., 2019?). Comparing pollinator population and community dynamics at large scales, i.e., between replicated urban landscapes, could provide insight into whether hypothesized urban habitat drivers significantly effect pollinators at large spatial and temporal scales.

We currently lack studies that make among-city comparisons, likely because of the logistical difficulties associated with collecting standardized data in a replicated format required to perform large-scale cross landscape comparisons. The proliferation of citizen science data and digitized museum records, present an opportunity to make inferences across large spatial and temporal scales. Occupancy modeling approaches that account for spatial and temporal variation in detection ability provide a framework for inferring population trends (Mackenzie). Occupancy of a species provides a proxy for population dynamics (Mackenzie), and whether implicitly (or explicity stated) at a broader spatial scale might provide a proxy for metapopulation dynamics (whether the metapopulation as a whole is stable or declining, and what environmental factors contribute to this). Paired with modeling choices that account for the unstructured and opportunistic nature of these data, this approach has emerging potential for uncovering broad global change impacts (Altwegg et al., 2019; Kharouba et al., 2019; Guzman et al., 2021; Jackson et al., 2022; Shirey et al., 2022; Davis et al, 2022)

**1b) Goals**

Goal: To identify how land use in cities impacts long-term dynamics of pollinator populations across urban landscapes. Using a set of data extracted from (briefly describe), I will examine the effect of differences in land use among urban areas on pollinator (occupancy? Diversity?) using an occupancy modeling approach (list a source REF). I focus on data available for pollinators and land use within urban areas in the state of California, which allows me to reduce unaccounted landscape heterogeneity, and control for differences in environmental/land use data. I also focus on comparisons among urban landscapes within a contemporary era (i.e., at least after year 2000), given availability of NHC data and quality environmental data, and also to avoid the assumption that urban landscapes were comparable in a pre-modern era. The taxonomic scope of this project will include two important pollinator taxonomic groups: hover flies (Diptera – family Syrphidae) and bumble bees (*Bombus*).

**1c) Hypotheses and predictions**

H1: Because greenspace is associated with nesting habitat (non-impervious surface for ground nesting as well as vegetative materials for aboveground nesting) and potentially with increased food supply (vegetation typically includes plants that offer pollen and nectar rewards), increased greenspace in the urban landscape positively impacts the stability of pollinator metapopulations.

Prediction: After accounting for a variety of urban characteristics (population density, area of the landscape site, and spatial autocorrelation between landscape sites) as well as heterogeneity in the detection probability of pollinators across space, time and taxa, the equilibrium **occupancy rates of pollinator species among urban landscapes will be positively associated with average NDVI** (a metric of proportion (?) green vegetation cover) in the landscape. [Note that this can be said very simply – urban areas that devote more space to plant cover – a predictor of available green space - are predicted to have higher pollinator biodiversity – it might be worth looking over the entire intro to be sure that is easily plucked out by the reader. Also, your intro suggests connectivity might also be important – can you add that to your analysis?]

* 1. **Data collection:**

Mention sensitivity analyses for data collection

* 1. **Statistical Analysis – a multispecies, integrated abundance-occupancy model:**

I use a multi-species (partially-pooling information about ecological and observation processes among species), integrated (incorporating different types of observed data that arise from a shared underlying ecological process), hierarchical (explicitly partitioning a latent ecological process and a conditional observation process) abundance-occupancy (explicitly linking abundance and occupancy) model to test whether hypothesized urban habitat features that vary among cities have significant impacts on urban pollinator populations.

***Ecological Process:***

I consider occupancy (*Z*) of a species (*i*) at a site (*j*)in a time interval (*k*)to be the outcome of a Bernoulli trial with probability of success equal to :

(1)

Correspondingly, I represent the latent abundance state (*N*) using a hurdle model, where is dependent on the Bernoulli process that drives the latent occupancy state . Specifically,is drawn from a truncated count distribution when a site is occupied (= 1). The probability of an abundance of 0, then, can be expressed as the complement of the probability of occupancy, and the probability of any abundance greater than 0 can be expressed as the product of the probability of occupancy and the probability of a discrete abundance state generated from a truncated count distribution (as in Welsh et al., 1996; Williams et al., 2017):

(2)

In this case, I use a negative binomial distribution as the count distribution, defined by a mean (the linear predictor , which includes ecological covariates expected to effect the abundance state) and a dispersion parameter (). However, a Poisson distribution could be used with the advantage of avoiding estimation of the additional parameter if the mean and variance of the counts are taken to be equivalent. Note, that in contrast to a zero-inflated count model (e.g., Smith et al. 2012), under this hurdle model ; a site that is occupied must have an abundance >= 1 and only the Bernoulli distribution with probability can produce abundances of 0.

Here, I explicitly link abundance and occupancy in the ecological model framework. There is a fundamental (mathematical and ecological) link between these two ecological quantities that is supported with empirical evidence, i.e., that abundance and occupancy rates are positively correlated (Gaston 1996; Welsh et al., 1996; Holt et al., 2002; Smith et al., 2012), including for insects (Sileshi et al., 2009). The advantage of explicitly linking abundance and occupancy includes improved computational performance and refined estimates for abundance (Smith et al., 2012). With the datasets utilized here, which include one dataset consisting of imperfectly detected abundance counts (citizen science records) and another dataset consisting of imperfectly detected presence/absence data (museum records), an explicit link between abundance and occupancy embedded in a hierarchical model allows for integrated estimation of the shared parameters for abundance and occupancy and reciprocal estimation of detection parameters. To link abundance and occupancy, and where is a global occupancy intercept and is the effect of species-, site-, and time-specific expected abundance on the occupancy rate, I define as:

(3)

With the addition of only two extra parameters, this definition of , allows flexibility in occupancy rates among species, sites, and time intervals (Smith et al., 2012).

***Observation Process:***

Detection biases in museum records and citizen science observations are likely to arise due to increased collection frequency in recent years, in areas with high human presence, or due more frequent collection of species that are charismatic, easy to identify, or are the target of conservation efforts/scientific research (Jackson et al., 2022; Shirey et al., 2022). I infer separate detection processes for citizen science record and museum record observation using a hierarchical model structure, allowing for full flexibility between the observation processes (as opposed to e.g., assuming that the detection rate of a species by citizen science methods is related to its detection rate in the museum records, or that the two processes change similarly through time). With this structure the variation among outcomes of temporally replicated data collection events are then used to infer the observation process while variation for each species between across sites and across time intervals are used to infer an underlying ecological process (driving the abundance and occupancy states) (Nichols, 2004; Kery & Royle 2016).

I consider the individual-level detection of abundance by citizen science records () to be the outcome of binomial trials, where , is the latent abundance state, with a probability (detection rate) , where is a linear predictor including covariates that are expected to affect the citizen science detection process. The variation in counts across visits (*l*) is used to infer the detection probability.

(4)

Similarly, I consider that the species-level detection of occupancy by museum records () to be binomially-distributed, but where is the number of surveys (years) in the time interval in which the species was detected at a site. Further, the likelihood of is here predicated on *n* trials, where indicates the number of years in which a community-wide survey occurred at the site during the time interval (and thus the maximum number of times that a species could be detected at that site during that time interval). The detection rate, , is a linear predictor including covariates that are expected to affect the museum detection process.

(5)

Failing to consider whether species could have been sampled (scaling based on the number of community sampling events) can cause underestimates of detection rates and consequently bias estimates of abundance (Guzman et al., 2021). Community sampling events each site in each year within each time interval were inferred prior to running the analysis. Specifically, if records from 2 or more species were collected from a site in a single year from the same institution, I designated that a community sampling event had occurred. A more conservative approach might be to group records by individual collector names rather than by institution, however, by visual review of the museum record data it appears that collector names even within the same year are variable (punctuation, inclusion of first and middle names) and, further, the museum records appear to include many ‘bioblitz’ type sampling events where a team of researchers or volunteers surveyed the community at same site on the same day. Community sampling events were inferred separately for the two taxonomic groups (Syrphidae and *Bombus*).

Note that in contrast to a classic binomial N-mixture model (Kery & Royle, 2016), the underlying abundance state estimated under this framework is a relative rather than true measure. This model violates two key assumptions required to estimate true abundance. First, the size of the population must remain constant between visits. The insects considered here have an annual life cycle with death of the adult generation each winter, and so we expect at least some inter-annual abundance variation to occur, precluding the estimation of the true abundance. However, if inter-annual abundance fluctuation within a time interval (here considered to be a set of 3 years) is non-trivial and non-random with respect to our ecological covariates (I think they talk about this in the Murray et al., population ecology textbook), then this modelling approach could produce biased estimates of the drivers of urban pollinator abundance. This limitation is a key drawback to N-mixture modelling approaches that is untestable without auxiliary data (I think they talk about this in the Murray et al., population ecology textbook).

Additionally, our approach violates the assumption that all individuals are available for detection during each visit (each year). In truth, insect abundances at these large spatial scales are likely orders of magnitude larger than the maximum number of records captured by citizen scientists (i.e., millions or more individual insects existing on the landscape versus a few or dozens of records obtained per year). Here, we consider that the number of records observed of a species at a site in each year by citizen scientists arises from a constant ‘pre-thinning’ process that reduces the millions of insects on the landscape to a proportional maximum number of insects that could be observed. In practice, this means that as variation of a species at a site across years within a time period decreases (e.g., counts of 99, 100, and 101) then detection rate is likely to be high and the observed abundance close in size to this abstract pre-thinned relative abundance; and in contrast, as variation of a species at a site across years within a time period decreases (e.g., counts of 5, 25, and 75) then detection rate is likely to be low and the observed abundances much lower in size to this abstract pre-thinned relative abundance. If the proportionality between true abundance and maximum observable abundance is constant, then variation in detection that obscures the ecological process should be reliably captured by the observation process covariates that allow and account for variation in sampling efforts.

The same model, including the same linear predictors for the ecological process and for the observation processes were fit independently for the two taxonomic groups, Syrphidae and *Bombus*.

***Linear predictor for ecological process:***

Where:

is a global, grand-mean occupancy intercept.

, and represents a species-specific intercept.

, and represents a site-specific intercept.

; where is the mean effect of impervious surface across species and describes the variation across species.

is the (scaled) proportion of impervious surface within the urban site *j*. (see data details for how this information was collected and how it was transposed onto the site given the site grid parameters specified in the run\_model.R program).

, where is the mean effect of the plant cover metric across species and describes the variation across species.

is the (scaled) plant cover metric of the urban site *j*. (see data details for how this information was collected and how it was transposed onto the site given the site grid parameters specified in the run\_model.R program).

is a fixed effect of spatial extent of the site. Some sites are not completely overlapping with a) land and b) the administrative area from which NHC data were drawn. A neutral process would suggest lower likelihood of occurrence and abundance in smaller sites (ref).

is the (scaled) spatial area of the site *j*. (see data details for how this information was collected and how it was transposed onto the site given the site grid parameters specified in the run\_model.R program).

***Linear predictors for observation processes:***

The log-odds of detection of an individual occurrence from a pool of available occurrences (assumed to be proportional to the true abundance of insects on the landscape) by the citizen science detection process is defined by the linear predictor:

*Where:*

is a global, grand-mean occupancy intercept.

, and represents a species-specific intercept.

, and represents a site-specific intercept.

is a fixed effect of population density on detection rate (i.e., more people means potentially more iNaturalist contributors).

is the (scaled) average human population density within the urban site *j*. (see data details for how this information was collected and how it was transposed onto the site given the site grid parameters specified in the run\_model.R program).

is a fixed effect of time on detection rate; and

is the temporal interval during which an observation was/was not made.

The log-odds of detection of one or more records (species-level detection versus non-detection) by the museum records sampling process is defined by the linear predictor:

is a global, grand-mean occupancy intercept.

, and represents a species-specific intercept.

, and represents a site-specific intercept.

is a fixed effect of population density on detection rate (i.e., more people means potentially more iNaturalist contributors).

is the (scaled) average human population density within the urban site *j*. (see data details for how this information was collected and how it was transposed onto the site given the site grid parameters specified in the run\_model.R program).

is a fixed effect of time on detection rate; and

is the temporal interval during which an observation was/was not made.

***Range restriction:***

At the large spatial scale considered here, not all sites are within the distributional range of all species, i.e., some species cannot occur at some sites. Failing to restrict the analysis for each species to sites within the range of each species can bias the parameter estimates for the ecological and observation processes (Guzman et al., 2021; Shirey et al., 2022).

To restrict the analysis to sites within the range of each species, I drew a convex-hull spatial polygon around all occurrence records from the year 2000 to present (library ‘sf’ in R), and determined which sites overlap with each polygon (as in Shirey et al., 2022). Using the intersections, I constructed a binary indicator array of sites within range versus sites outside of range for each species (held constant across time intervals) that is used to forbid out-of-range species by site combinations from contributing to parameter likelihood.

***Software and Model Implementation:***

Analyses were conducted using STAN version 2.21.0 (ref), implemented through the R interfacing package “rstan” using R version ().

In brief, the posterior likelihood of the parameters given the data and the priors from each iteration of the MCMC algorithm is defined by marginalizing across different possible latent states. If a species was observed at a site (within the range) in a time interval by either NHC data set, then it is considered to both occupy the site and have an abundance greater than 0. The likelihood of the parameters across a vector of abundance from the maximum observed abundance to an abundance of K is calculated. If, however, a species was never observed at a site (within the range) in a time interval by either NHC data set, then the species could either a) not occupy the site and have an abundance of 0, or b) be present at a site with some abundance greater than 0, but was never detected. The likelihood of the parameters across a vector of abundance from 0 to K is calculated, and the log probability of the parameters is summed.

In either case the value of K is generated before fitting the model, where I defined K as:

The addition of 5 allows for a possibility for an latent relative abundance greater than 0 despite any detections, and the multiplication by 5 all sets the cap on the latent relative abundance search as 5 times greater than the max observed count plus 5. K should be carefully chosen and evaluated, given that values of K that are lower than the latent relative abundance will truncate the estimates for the linear predictor as it will hold the parameter search back from estimating a larger latent abundance with lower detection rates, while values of K far greater than the latent abundance can significantly decrease computational efficiency and model run times (Kery & Schaub, 2012).

Data and code with instructions for fitting the model are available at: https://github.com/jensculrich/occupancy\_model\_for\_urban\_NHC\_records

**Prior Distributions:**

Most priors were chosen to be weakly informative. Normal(0, 2) priors were used for parameter means for the abundance linear predictor, intended to hold the posterior distributions of the parameters close to 0 (no effect of parameters on detection or abundance-occupancy) unless the data provide convincing evidence to the contrary. Normal(0, 2) priors were used for parameter means for the detection log-odds linear predictor, intended to hold the posterior distributions of the parameters close to 0 (parameters do not shift detection rates for individuals or species, respectively, above or below 50/50 random chance) unless the data provide convincing evidence to the contrary.

Weakly-informative half-Normal(0, 1) priors were used for all variance parameters, which encourages stronger pooling within clusters (e.g., species abundance responds fairly similarly to a covariate of impervious surface cover), unless the data provide convincing evidence to the contrary.

A strongly-informative prior is used for (effect of abundance on occupancy). Specifically, I use a half-normal prior (lower bound at 0) with a mean of 0 and standard deviation of 1. The lower bound on the prior is based on the theoretical and empirically supported relationship between abundance and occupancy (ref), forcing the abundance-occupancy relationship to be positive. This strong prior choice was implemented to assist with model fit and identifiability of the model given the large number of parameters to be estimated from the data.

The priors (could be) widened by a factor of say 2 or 3 and the results rexamined to examine whether the priors have strongly influenced the results.

**Simulation Checks:**

Prior to application on real data, the model was fit to a dataset with a similar sample size and that was simulated from known parameter distributions. This allowed us to confirm that the model accurately estimates true parameter values, given that the assumptions of the model are met.

**Model Diagnostics:**

The model outputs were assessed to confirm that there is sufficient mixing of the MCMC chains (all Rhat values < 1.1) with no/low divergent transitions, and an effective sample size of >100 for all parameters. The model is set to run for n=1,000 iterations with ½ n discarded as burnin, with an adapt delta value of 0.8. These values can all be upped for the real model run, but just leaving short for now while I work out the model application.

**Posterior Predictive Checks:**

A chi-squared discrepancy test (Kery & Schaub, 2012) was used to evaluate the goodness-of-fit of the model. The discrepancy test compares the squared distances between observed counts and detection/non-detection created by a random number generation and the parameters in each step of the MCMC algorithm versus the expected values for the counts given the parameters in each step of the MCMC algorithm to the squared distances between real NHC citizen science counts and museum record detections/non-detections versus the expected values for the counts given the parameters in each step of the MCMC algorithm.