Replication of “Historical redlining is associated with increasing geographical disparities in bird biodiversity sampling in the United States”

Martin Bulla[1,✉] and Peter Mikula[2]
(bullab, Faculty of Environmental Science, Czech University of Life Sciences Prague)

—

[1] bullam@fzp.czu.cz
[2] petomikula158@gmail.com

We were supported by the Research Excellence in Environmental Sciences grant from the Faculty of Environmental Sciences CZU (REES 03 to MB) and declare no conflict of interest.

—

Abstract

Ellis-Soto et al. (2023, Nature Human Behaviour) investigated whether the density and completeness of bird biodiversity sampling from citizen science observations across US cities covary with 1930s neighbourhood classifications based on perceived mortgage investment risk, a practice known as “redlining”. They claimed that worst-rated neighbourhoods were the most under-sampled urban areas for bird biodiversity and that such disparity in sampling increased between 2000 and 2020 by 35.6%.

We were initially unable to reproduce the data generation and analyses with the deposited code and data, but reproduced most reported findings after the authors provided missing data and we corrected several coding issues. However, the code underlying the spatial results (Fig. 1) and the key claim of a 35.6% temporal increase were absent. After correcting a major data-coding error (unintentional data multiplication), we recreated the temporal trends, including Fig. 4, in a manner consistent with the underlying true yearly data. We also demonstrate that Fig. 4 and the original claim of 35.6% increase arise from annual aggregates of sampling density that implicitly treat thousands of polygons as a single annual observation, precluding any modelling of spatial or temporal non-independence.

Despite model misspecification and annual data aggregation in the original analyses, our alternative analytical choices (a) reproduced the reported disparity among HOLC grades with effect sizes that differ in magnitude but are consistent in direction, and (b) revealed substantially more complex temporal trends. Specifically, the relative disparity between the worst- and best-rated neighbourhoods varied non-linearly over time and exceeded 350% by 2020. In contrast, absolute differences remain negligible until ~2010, rise to only ~5 observations per km² per year by ~2017, and reach at most ~25 observations by 2020 – small absolute magnitudes despite large relative changes.

To conclude, large relative disparities between the best- and worst-rated neighbourhoods do not imply large absolute differences in sampling density. Overall, disparity changes exceed ~200% between 2000 and 2020, which is inconsistent with the original authors’ claim of a 35.6% increase. The post-2010 rise coincides with the surge in smartphone-driven community science. The plateau around 2015 plausibly reflects broader smartphone accessibility. The sharp rise in 2020 aligns with COVID-19 restrictions, which markedly boosted urban green-space use and citizen science participation.

Introduction

Ellis-Soto, Chapman, and Locke (2023)[1] examined whether the extent of bird biodiversity sampling in US city neighbourhoods correlates with neighbourhoods' historical 1930's categorization according to perceived safety for real estate investment. This practice is known as redlining and was operationalised using Home Owners' Loan Corporation (HOLC) grades (A–D, with A being perceived as the safest and D as the least safe).

Specifically, the authors claim to have investigated:

> "*after controlling for greenspace and climate… (1) how does sampling effort of bird biodiversity [neighbourhood sampled (yes or no), sampling density and completeness] vary with socioeconomic conditions (historic housing segregation policies, such as HOLC-graded neighbourhoods), biophysical predictors across urban environments and by data source; (2) whether the least-sampled areas for bird biodiversity (biodiversity coldspots hereafter) are consistently located in regions that were segregated by redlining; and (3) whether when considering temporal trends, biodiversity sampling is becoming more even or uneven over time in the age of increasing digital and citizen-science data collection?*"[1] (main text, p. 1870).

Based on their results, the authors concluded that "*historically redlined neighbourhoods remain the most undersampled urban areas for bird biodiversity today, potentially impacting conservation priorities and propagating urban environmental inequities*"[1] (main text, p. 1870). They also claim that this sampling disparity *"increased by 35.6% over the past 20 years"*1 (abstract, p. 1869).

Replication

We had difficulties linking the reported results (particularly the model outputs) to those generated by the authors' code. This was due to ambiguous description of the statistical methods, and insufficient details in the README file and script annotations (see our supporting information and its Table S1[2]).

Computational reproducibility

Using the deposited code, data and instructions, we were initially unable to reproduce any findings because the code referred to essential datasets that were not deposited (see our Table 1 and the *Computation reproducibility* section of the supporting information[2]).

Table 1 | Provided data and code, and reproducibility of the results.

		Fully	Partial	No
Provided	Raw data		x	
	Cleaning code		x	
	Analysis data		x	
	Analysis code		x	
Reproducable	From raw data			x
	From analysis data		o	x

x - initially, o - after further data provided from the authors and our adjustments to the code.

After the authors shared the missing data, we were still unable to fully reproduce the data-generation process because (i) the GBIF dataset cited in the paper (https://doi.org/10.15468/dd.ha9ksv) could not be accessed, and (ii) to recreate it, the workflow required logging into the GBIF platform with credentials, a step we were unable to execute. Despite these limitations, we successfully reproduced most main-text and supplementary outputs, including Figs. 2–3, S1-S2, and Tables S1–S3, after correcting minor code issues and relying on the already prepared, deposited datasets (see also our Table S1). However, we could not reproduce Fig. 1 and the key claim that sampling disparity increased by 35.6% over time, as the authors did not describe how these outputs were generated, and we could not identify any corresponding code.

We only partially reproduced Table S2. Specifically, we reproduced the first panel, labelled as ‘log(sampling density)’, but displaying untransformed sampling density (see our Table S2). The second panel, also labelled as log(sampling density), could not be reproduced. Furthermore, according to the table caption, three panels were expected: sampling presence, sampling density, and sampling completeness. However, it remains unclear which sampling or completeness models should have been depicted.

Initially, we were also unable to fully reproduce (i) Table 1 because the model shown was not the best supported according to the Akaike Information Criterion model selection implemented in the script, and (ii) Table S4 because the model description did not match the script and the underlying dataset was accidentally multiplied (our Table S3). The latter issue also obscured the reproducibility of Fig. 4 (see the next section).

We identified a **major coding error** in the script preparing the temporal trend data, which had multiplied the dataset (our Figs. S1–S2). Hence, Fig. 4, Table S4, and likely the reported 35.6% increase in disparity did not represent the underlying data (see our Fig. S2 and Table S3).

Having corrected this error, we recreated the temporal trends, including Fig. 4 and Table S4, in a manner consistent with the underlying true yearly data (see our Figs. S2–S3 and Tables S3–S4[2]). In doing so, we found out that Fig. 4 and likely the original claim of 35.6% were based on coarse annual aggregates of sampling density by HOLC grade. This aggregation ignores variation in observations at the polygon level and prevents modelling of spatial or temporal non-independence. This renders the aggregated trends statistically fragile and difficult to interpret (see our Figs. 4 and S4).

Robustness reproduction

Some authors’ models contained over-specified and/or misspecified random effects structures and were missing intended control variables (see our supporting information[2]). In addition, the authors used the Akaike Information Criterion to compare models differing in both fixed and random effects, a practice that is generally discouraged and unnecessary in the present context[3–5].

HOLC-grade differences - After addressing the above issues by re-specifying the models, our estimates consistently reproduced the qualitative pattern of reduced sampling in lower HOLC grades across response variables, model structures (from minimal to fully controlled), and error distributions. However, absolute differences were sensitive to data transformation and model specification (our Figs. 1–3).

For the probability that a neighbourhood was sampled, binomial and Gaussian models yielded closely aligned results (our Fig. 1), confirming robustness to link-function choice and distributional assumptions, despite non-normal data (cf.[6]). Including neighbourhood area alone reduced the apparent A–D difference, indicating that part of the original HOLC grade effect reflected variation in polygon size rather than sampling bias. However, adding additional covariates (normalised difference vegetation index, proportion of protected area, population density and climate), strengthened the A–D contrast again, suggesting that ecological context masked some social patterns in simpler models. Overall, these controls stabilised the estimates while preserving their direction and approximate magnitude. In contrast, our re-specified

Gaussian models yield more extreme A–D contrasts (farther from zero) than the authors' original results (our Fig. 1), suggesting that the inclusion of relevant controls reveals stronger HOLC grade differences that were partially masked in simpler models.

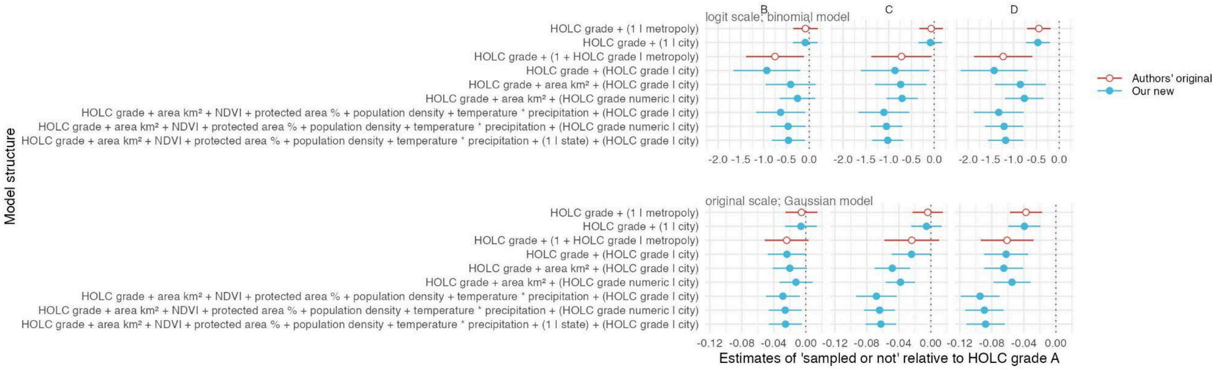


Figure 1 | Differences in estimated sampling presence between HOLC grades across modelling approaches. Points represent fixed-effect contrasts for HOLC grades B-D (columns), relative to HOLC grade A, in the probability that a HOLC polygon was ever sampled (binary "sampled or not" outcome). Horizontal lines are 95% Wald confidence intervals, and the vertical dashed lines indicate zero difference. Red open circles indicate models specified by the original authors; blue filled circles those specified by us. The y-axis lists model structure: terms outside parentheses are fixed effects (* denotes interactions), terms inside parentheses denote random effects (variables left of | are random slopes, variables right of | are random intercepts). The **top row** shows binomial logit-link models of sampling presence, the **bottom row** Gaussian identity-link models fitted to the same binary outcome on the original 0–1 scale. All models are based on n = 9,847 HOLC polygons (neighbourhoods).

The authors modelled log(sampling density) using Gaussian models, either excluding zeros or adding a small constant. Our modelling of both these variants reproduced the qualitative pattern of reduced sampling in lower HOLC grades. When re-analysed as count data with an offset for polygon area and a negative-binomial family, the direction of effects remained consistent, but more stable across model specifications (our Fig. 2). This stability reflects that the count-offset formulation properly accounts for unequal polygon areas, includes zeros without arbitrary transformation, and models overdispersion explicitly rather than absorbing it into residual variance. Across all variants, the conservative random-effects structure (city) yielded comparable estimates while better controlling for local clustering of observations.

For the completeness of sampling, our replication quantifies a larger disparity for A-D than authors (our Fig. 3), and demonstrates that robustness improves with adequate random-effect specification.

Temporal trends in HOLC-grade differences - To assess changes over time, the authors aggregated all observations within each year and HOLC grade. They then calculated the relative difference between the best- (A) and worst-rated (D) neighbourhoods, and estimated the percentage change in such disparity between 2000 and 2020. However, such type of endpoint comparison implicitly assumes monotonic change and stable variance, neither of which holds for these data.

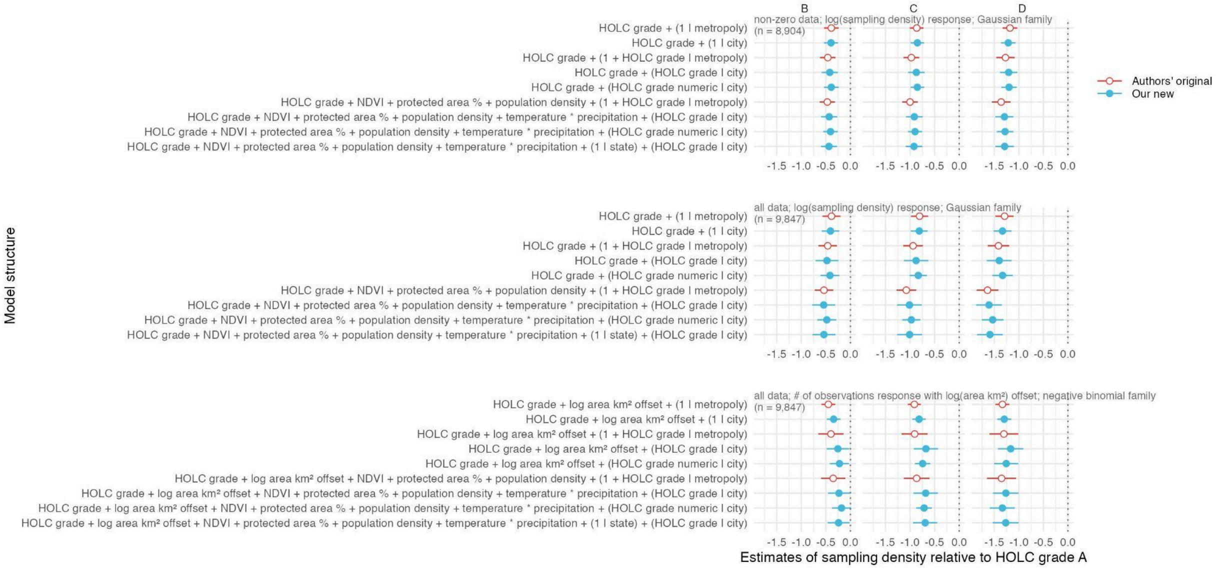


Figure 2 | Differences in estimated sampling density between HOLC grades across modelling approaches. Points represent fixed-effect contrasts for HOLC grades B-D (columns), relative to HOLC grade A, in sampling density of bird observations (per km^2), while horizontal lines are 95% Wald confidence intervals, expressed on the model's link scale. The vertical dashed lines indicate zero difference. The **top row** shows Gaussian identity-link models fitted to log-transformed sampling density using only polygons with non-zero sampling density; the **middle row** shows Gaussian identity-link models fitted to all polygons (including zeros) by adding a small data-derived constant to the sampling density to allow log transformation (log(sampling density + 0.125)); the **bottom row** shows negative-binomial log-link models of bird observation counts with an log(area (km^2)) offset, modelling counts per unit area and thus representing differences in sampling density. Sample sizes for each row are given in the subtitles. The vertical dashed line indicates zero difference. Red open circles indicate models specified by the original authors; blue filled circles those specified by us. The y-axis lists model structure: terms outside parentheses are fixed effects (* denotes interactions), terms inside parentheses denote random effects (variables left of | are random slopes, variables right of | are random intercepts).

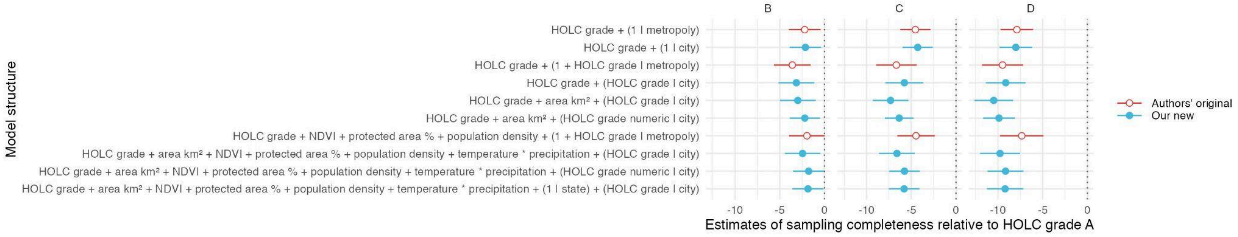


Figure 3 | Differences in estimated sampling completeness between HOLC grades across modelling approaches. Points represent fixed-effect contrasts for HOLC grades B-D (columns), relative to HOLC grade A, in completeness of sampling (index), while horizontal lines are 95% confidence intervals. The vertical dashed line indicates zero difference. Red open circles indicate models specified by the original authors; blue filled circles those specified by us. The y-axis lists model structure: terms outside parentheses are fixed effects (* denotes interactions), terms inside parentheses denote random effects (variables left of | are random slopes, variables right of | are random intercepts). All models are based on n = 7,187 HOLC polygons (neighbourhoods).

Visualising the A–D disparity in yearly aggregated sampling density through time reveals a strongly nonlinear dynamic (our Figs. 4 and S4), with stable or decreasing disparity in the early 2000s, followed by a rapid increase around 2010 and subsequent periods of stagnation or decline. The relevance of comparing disparity only between the years 2000 and 2020 is therefore questionable. Notably, the disparity in the proportion of sampled neighbourhoods has decreased since ~2008, and absolute differences were close to zero for most of the studied period (i.e. until ~2008; our Fig. S4). Note that the raw sampling densities of A and D polygons were nearly indistinguishable until ~2010 (Figs. 4 and S4). Furthermore, disparity trajectories depend on the aggregation method (our Figs. 4 and S4; see our supporting information for details[2]) and fitting linear models to such aggregates is likely to be misleading (our Table S4 and Figs. S5–S7).

Aggregating data by year and HOLC grade fails to account for polygon-level heterogeneity (our Figures 5–6, S4–S5) and ignores the non-independence of data points in space and time, which biases the results. We thus analysed the raw, yearly, polygon-level data (i.e. the number of observations per sampling polygon in each year) using models that explicitly accounted for such dependence.

We found that only the A–D contrast showed a statistically supported difference in slopes, and even this effect was weak (our Fig. 6). The slope estimates were nearly identical across grades (particularly for B–D) and became statistically uncertain when analyses were restricted to the 2010–2020 period (Fig. 6). This indicates that apparent differences in slopes are partly driven by misfit to nonlinear temporal patterns rather than sustained divergence. Although these models (our Figs. 6 and S9) provide appropriate alternatives to the authors' misspecified analyses, they impose linear relationships. Our results, however, suggest that temporal dynamics in sampling disparity are more complex (our Figs. 4 and S4–S6), motivating the use of flexible smooth models.

To capture these nonlinear dynamics, we analysed temporal trends using smooth models (see our supporting information[2]). These revealed a constant relative disparity of ~100% in the first five years (2000–2005), which increased to ~250% by 2010 and remained stable or declined thereafter. It then rose sharply after 2018 reaching ~350% by 2020.

Overall, relative disparity increased by ~200% between 2000 and 2020, which contrasts with the 35.6% increase reported by the authors. However, these relative disparities correspond to small absolute differences: near zero until ~2010, ~5 observations per km^2 per year by ~2017, and at most ~25 observations by 2020 (our Fig. 7). Thus, although relative disparities appear visually dramatic, absolute differences remained small for most of the study period. This apparent inflation of relative disparities largely arises from relative scaling rather than large biological effects: when the sampling density in D-rated neighbourhoods is extremely low (<0.1), even small absolute differences yield large percentages. Also, note that the disparities are neither linear nor monotonic (our Fig. 4 and 7).

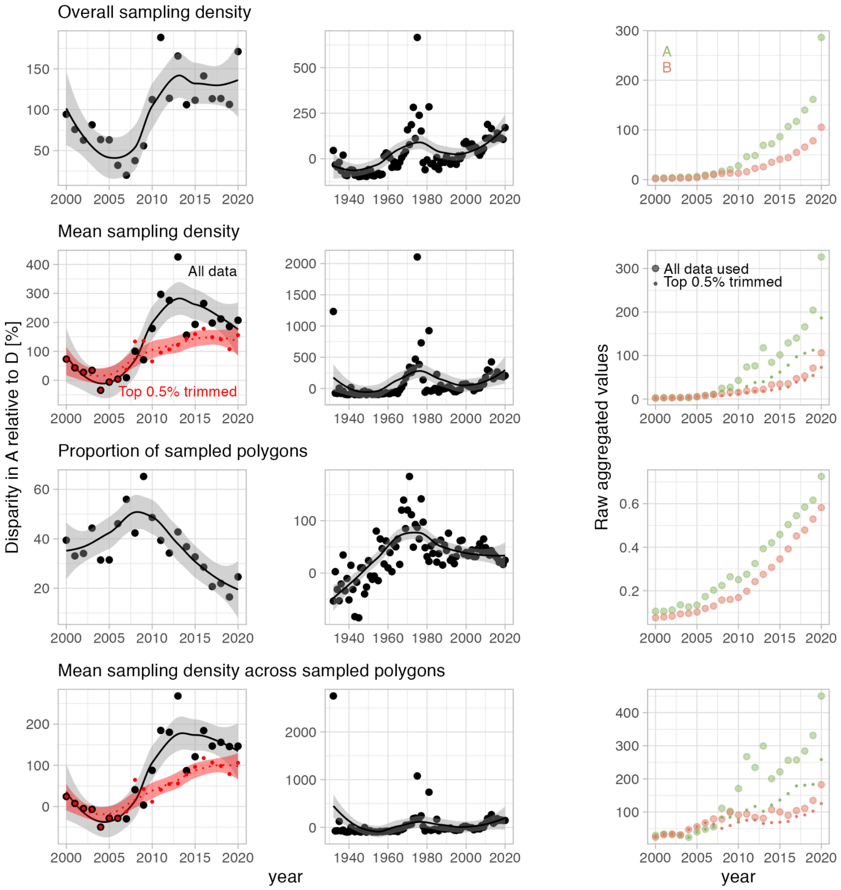
In sum, the authors' finding of 35.6% change in disparity only compares endpoints (2000 vs. 2020), which implicitly treats disparity as a single linear change. However, our analyses reveal that temporal trajectories in disparities are strongly nonlinear and fluctuate substantially, with disparities exceeding 200% over the same period. The rapid increase around 2010 coincides with the introduction of smartphones[7] and the expansion of citizen-science platforms such as eBird and iNaturalist[8,9]. Subsequent levelling off may reflect broader smartphone accessibility[10,11]. The sharp rise in sampling and disparity in 2020 aligns with COVID-19-related increases in local greenspace use.

Conclusions

To summarise, computational reproducibility of the key analyses was not possible with the deposited code and data, but was achieved for most results once the authors provided the missing data and several minor coding issues were resolved. Correcting a major coding error enabled us to recreate the temporal trends, including Fig. 4, in a manner that was consistent with the underlying true data.

Despite issues with the original model specification and data aggregation, our alternative analytical approach reproduced the reported HOLC-grade differences while revealing strongly nonlinear temporal dynamics. The relative disparity between the best- and worst-rated neighbourhoods increased substantially over time, reaching ~350% by 2020 (an overall change in disparity of ~200% between 2000 and 2020, compared with the originally reported 35.6%). However, these large relative changes corresponded to small absolute differences, which remained negligible until ~2017 and reached at most ~25 observations km^{-2} yr^{-1} by 2020.

Supporting material, including the code and data generating the outputs is freely available at https://martinbulla.github.io/I4R_Nat_Hum_Beh/[2].



Overall sampling density

Mean sampling density

Proportion of sampled polygons

Mean sampling density across sampled polygons

Figure 4 | Change in relative disparity in sampling density between HOLC grade A and D over time. Left two columns, each point represents relative difference (%) in sampling density of A given D (with D being a baseline) based on overall sampling density (i.e. sum of all A or D observation divided by the total area of A or D; **1st row**), mean sampling density per HOLC grade and year (**2nd row**), proportion of sampled polygons (**3rd row**) and mean sampling density across sampled polygons (i.e. excluding non-sampled ones; **bottom row**). The **right column** shows the actual values for A and D HOLC grades. Dots represent yearly values (for all data: large dots; for data with top 0.5% observations trimmed: small dots). Lines represent local regression non-parametric smoothing and shaded areas 95% confidence intervals. Colour in the right column indicates all data (black) or data with top 0.5% trimmed (red), in the left column the HOLC grade category (A in green, D in red). The top row represents the aggregation likely used by the authors to support their claim, the other rows the aggregation done by us.

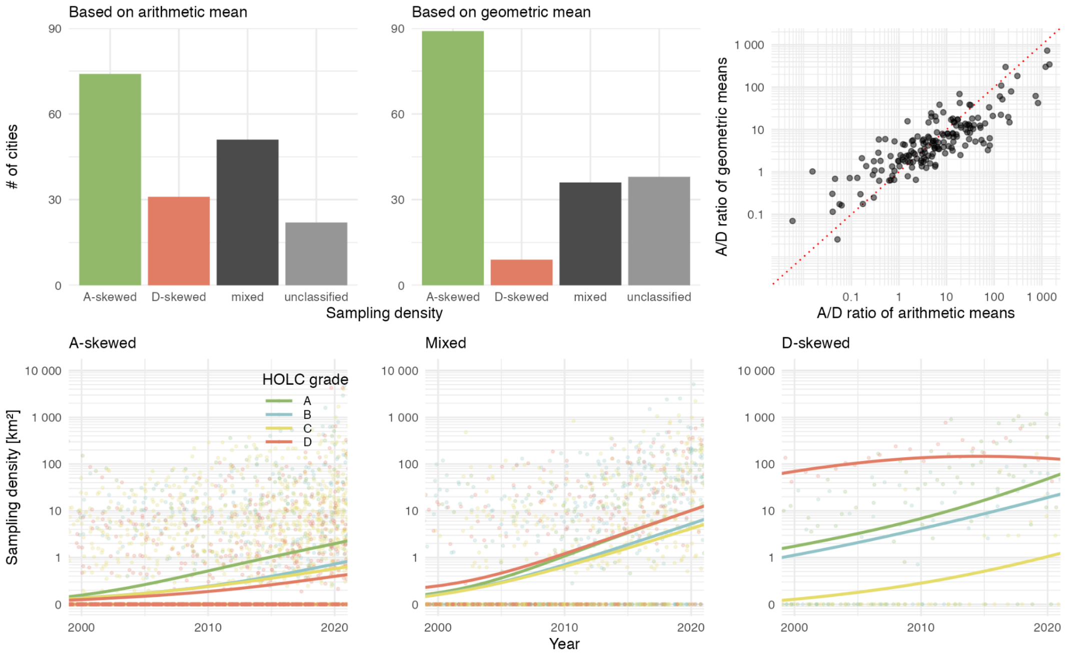


Figure 5 | **City-level variation in sampling density and HOLC grade skew. Top row**, number of cities classified as A-skewed, D-skewed, mixed, or unclassified based on arithmetic-mean (left) or geometric-mean (middle) sampling density ratios between HOLC A and D polygons (for details see supporting information[2]). Comparison of city-level A/D ratios from geometric vs arithmetic means (right); points above the 1:1 line indicate cities where arithmetic means underestimate A-skew (typically due to strong D-grade hotspots raising the arithmetic mean for D), while points below the line indicate cities where arithmetic means overestimate A-skew (typically due to strong A-grade hotspots raising the arithmetic mean for A). Arithmetic means are highly sensitive to rare but extreme polygons ("hotspots"), and therefore reflect occasional survey campaigns more than the underlying spatial structure. Geometric means minimise outliers and capture the "typical" polygon in the "typical" year. The wide scatter around the 1:1 line shows that no single aggregation metric yields a stable classification, cities frequently flip between A-skewed, mixed, and D-skewed depending on whether hotspots are emphasised (arithmetic) or down-weighted (geometric). **Bottom row**, representative example cities: A-skewed city (left), where A-grade polygons are consistently sampled more densely than D-grade polygons, mixed city (middle) with no persistent ordering between A and D grades across years, and D-skewed city (right), where D-grade polygons receive higher sampling density. Panels show raw polygon-level sampling densities (points) with local regression non-parametric smoothing trends per HOLC grade (solid lines). Cities were selected using a data-driven procedure based on the geometric mean A/D ratio (see supporting information[2]). These examples illustrate that within-city patterns vary substantially and help contextualize the aggregate national-level disparity trends shown in Figs. 4 and S4. The plots for each city are in Fig. S5.

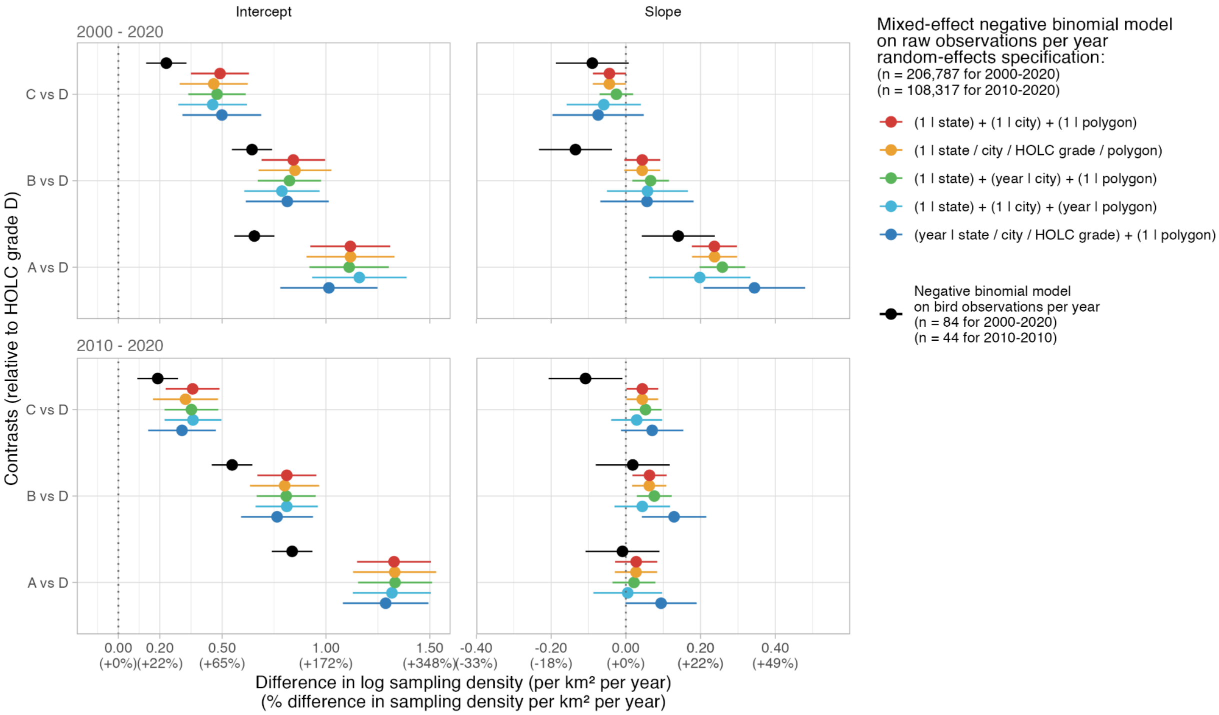


Figure 6 | Estimated differences in HOLC grade sampling density over time. Dots represent fixed-effect contrasts on the log scale, together with the implied percentage change in sampling density (observations per km^2 per year; relative to grade D), obtained by exponentiating those contrasts from negative binomial mixed models with log link and offset of log(area(km^2). Intercept panels show differences between each HOLC grade at the mean year, slope panels differences per standard deviation increase in year. Horizontal lines are 95% Wald confidence intervals. The vertical dashed line indicates zero difference. Colour indicates random-effects structures (variables left of | are random slopes, right of | random intercepts, and / indicates nesting). The **top row** contains estimates for a dataset spanning 2000-2020 (for the linear model n = 84 sum of observations per grade and year, for the mixed models n = 206,787 sum of observations per polygon and year); the **bottom row** contains estimates for a dataset from 2010-2020 (n = 44 and n = 108,317, respectively). For per grade estimates see Fig. S8.

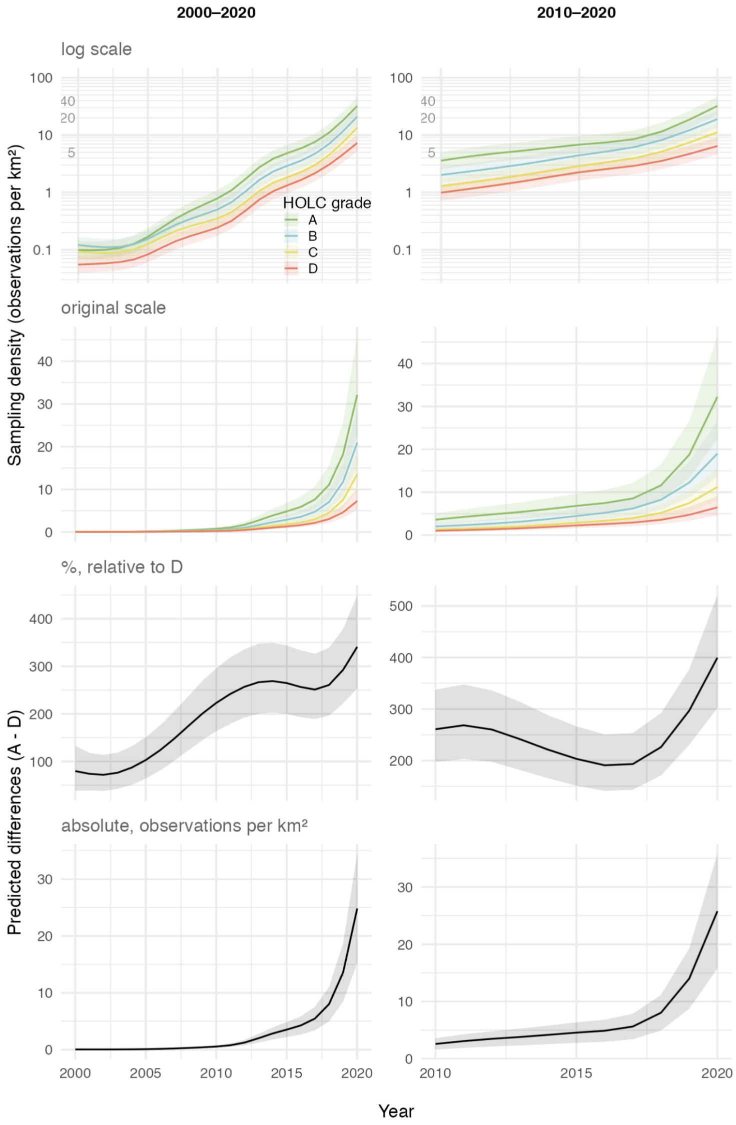


Figure 7 | Non-linear temporal changes in bird-sampling density by HOLC grade and disparity between grades A and D. Top two rows, population-level (marginal) predictions from a negative binomial generalised additive model (bam) with log link and centered log(area(km^2) offset, fitted to polygon-level counts for 2000–2020 (**left**) and 2010-2020 (**right**). The model included a smooth for year, grade-specific smooth deviations, and random effects for state, city, and polygon, and city-specific temporal slopes. Curves show predicted sampling density (observations per km^2) for each HOLC grade on a log scale (**1st row**) and original scale (**2nd row**). **Bottom two rows**, predicted differences between grades A and D from the same model fitted to observations from 2000-2020 (**left**) and 2010-2020 (**right**), expressed as percent difference relative to D (**3rd row**) and as absolute difference in observations per km^2 (**bottom row**). Relative disparity varies non-linearly over time and exceeds 350% by 2020, whereas absolute differences remain small (< ~25 observations per km^2), indicating that large proportional disparities do not translate into large absolute changes in sampling intensity. In **all panels**, shaded ribbons are 95% confidence intervals.

References

1. Ellis-Soto, D., Chapman, M. & Locke, D. H. Historical redlining is associated with increasing geographical disparities in bird biodiversity sampling in the United States. *Nat. Hum. Behav.* **7**, 1869–1877 (2023).
2. Bulla, M. & Mikula, P. Supporting information for Replication of “Historical redlining is associated with increasing geographical disparities in bird biodiversity sampling in the United States”. *GitHub* https://martinbulla.github.io/I4R_Nat_Hum_Beh/ (2025).
3. Zuur, A. F., Ieno, E. N., Walker, N., Saveliev, A. A. & Smith, G. M. *Mixed Effects Models and Extensions in Ecology with R*. (Springer, New York, NY, 2009). doi:10.1007/978-0-387-87458-6.
4. Greven, S. & Kneib, T. On the behaviour of marginal and conditional AIC in linear mixed models. *Biometrika* **97**, 773–789 (2010).
5. Vaida, F. & Blanchard, S. Conditional Akaike Information for Mixed-Effects Models. *Biometrika* **92**, 351–370 (2005).
6. Knief, U. & Forstmeier, W. Violating the normality assumption may be the lesser of two evils. *Behav. Res. Methods* **53**, 2576–2590 (2021).
7. August, T. *et al.* Emerging technologies for biological recording. *Biol. J. Linn. Soc.* **115**, 731–749 (2015).
8. eBird, T. eBird mobile app for iOS now available! - eBird. https://ebird.org/ebird/news/ebird_mobile_ios1 (2015).
9. eBird, T. Celebrating eBird’s 20th Anniversary - eBird. https://ebird.org/ebird/news/ebird-20th-anniversary (2022).
10. DeSilver, D. The falling price of a smartphone. *Pew Research Center* https://www.pewresearch.org/short-reads/2013/09/10/the-average-selling-price-of-a-smartphone/ (2013).
11. Smith, A. Chapter One: A Portrait of Smartphone Ownership. *Pew Research Center* https://www.pewresearch.org/internet/2015/04/01/chapter-one-a-portrait-of-smartphone-ownership/ (2015).