Data aggregation obscures temporal trends in bird sampling

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

—

[1] bullam@fzp.czu.cz
[2] mikulap@fzp.czu.cz
—

Ellis-Soto et al.[1] reported that disparity in bird-sampling density between U.S. city neighbourhoods rated as safe versus risky for real estate investment (a practice known as "redlining") increased by 35.6% between 2000 and 2020. We show that this reported trend arises from data aggregation and linear model misspecification. Using the original neighbourhood-level yearly data and mixed-effects models that account for spatial and temporal non-independence, we demonstrate that temporal disparities are non-linear and exceed 200% across the study period, while absolute differences remain small (near zero for most of the study period and reaching at most 25 observations per km² per year). These non-linearities temporally coincide with major shifts in citizen-science participation, including smartphone adoption and COVID-19-related increases in urban greenspace use.

The code underlying the authors' key claim of a 35.6% temporal increase in disparity between safe- (A) and risky-rated (D) U.S. neighbourhoods was missing. Most other analyses could be reconstructed once missing data and minor coding issues were resolved, and they were generally robust to analytical decisions (see Supporting information[2]).

Here, we demonstrate that the authors' temporal results arise from annual aggregates of sampling density, which implicitly treat thousands of neighbourhoods as a single annual observation, obscuring heterogeneity among neighbourhoods and precluding any modelling of spatial or temporal non-independence (our Fig. 1). We then show that accounting for such non-independence yields distinct, non-linear temporal trends in sampling disparity between safe- and risky-rated neighbourhoods (Fig. 2).

Consequences of annual aggregation

To assess changes over time, the authors aggregated all bird observations within each year by 1930's Home Owners' Loan Corporation (HOLC) grades (A–D). The authors then calculated the relative difference between safe- (A) and risky-rated (D) neighbourhoods and estimated the percentage change in such disparity between 2000 and 2020. However, this type of endpoint comparison implicitly assumes monotonic change and stable variance, neither of which holds for these data.

Visualising the A–D disparity in yearly aggregates of bird sampling density reveals a strongly non-linear dynamic (our Fig. 1), 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 relative A–D disparity in the proportion of sampled neighbourhoods has decreased since ~2008, and the raw sampling densities of A and D neighbourhoods were nearly indistinguishable until ~2010 (our Fig. 1).

Crucially, disparity trajectories depend on the aggregation method (our Fig. 1; see also our Supporting information2 for details), and fitting linear models to such aggregates may be misleading. Aggregating data by year and HOLC grade also fails to account for neighbourhood-level heterogeneity (our Extended Data Fig. 1–2) and ignores the non-independence of data points in space and time, thereby biasing the results. We thus analysed the raw yearly neighbourhood-level data (i.e. the number of observations per sampling polygon per year) using mixed-effects models that explicitly accounted for such dependence (see Supporting information[2] for model specification).

Only the A–D contrast showed a statistically clear difference in slopes[3], but even this effect was weak (Extended Data Fig. 3). Slope estimates were nearly identical across grades (particularly for B–D) and became statistically unclear when analyses were restricted to the 2010–2020 period (Extended Data Fig. 3), indicating that apparent differences in slopes across 2000–2020 are partly driven by model misfit rather than sustained divergence. Although our models provide appropriate alternatives to the authors' original analyses, the authors' analyses are not fully supported by the data structure (e.g. with respect to the random-effects structure) and also impose linear relationships. Our results instead suggest more complex temporal dynamics in sampling disparity (Fig. 1 and Extended Data Fig. 1–3), motivating the use of flexible smooth models (see Supporting information[2] for model specification)..

The smooth models revealed a constant relative disparity of ~100% in the first five years (2000–2005), which increased to ~250% by 2010 and then remained stable or declined. After 2018, disparity rose sharply, reaching ~350% by 2020 (Fig. 2). Overall, relative disparity increased by ~200% between 2000 and 2020, which contrasts with the 35.6% increase reported by the authors. However, these large relative disparities correspond to small absolute differences in sampling density (observations per km² per year): near zero until ~2010, about five by ~2017, and at most ~25 by 2020 (Fig. 2). Thus, although relative disparities appear visually dramatic, absolute differences remained small for most of the study period. This inflation of relative disparities largely reflects relative scaling on very small baselines rather than large absolute differences in sampling intensity: when sampling density in D-rated neighbourhoods is extremely low, even small absolute differences yield large percentages. The disparities are neither linear nor monotonic (our Fig. 2).

The rapid increase in sampling and disparity around 2010 coincides with the introduction of smartphones[4] and the expansion of major citizen-science platforms such as eBird and iNaturalist[5,6]. Subsequent levelling off may reflect broader smartphone accessibility[7,8], whereas the sharp rise in 2020 aligns with COVID-19-related increases in local greenspace use and public participation in citizen science[9,10]. Data beyond 2020 will help clarify whether recent changes represent transient fluctuations or sustained trends.

Conclusions

Despite issues with the original model specification and data aggregation, our alternative analytical approach reproduced the reported HOLC-grade differences while revealing strongly non-linear temporal dynamics in sampling disparity. Relative disparity between safe- and risky-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 per km² per year by 2020.

Data & code availability

Supporting information, including the code and data used to generate the results, are freely available at https://martinbulla.github.io/MA_NHB/[2].

Acknowledgements

This project was initiated as part of the "Promoting reproduction and replication at scale" partnership between *Nature Human Behaviour* and the Institute for Replication (*Editorial 2024*; https://doi.org/10.1038/s41562-024-01818-7). We thank the Institute for Replication for guidance and support throughout the replication process, and the original authors for valuable feedback. We were supported by the Research Excellence in Environmental Sciences grant from the Faculty of Environmental Sciences, Czech University of Life Sciences (REES 03 to MB). We thank Anička, Bara and Majlen for their patience and support.

Competing interests

We declare no conflict of interest.

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 "Data aggregation blurs inferred temporal trends in bird sampling". *GitHub* https://martinbulla.github.io/MA_NHB/ (2026).
3. Dushoff, J., Kain, M. P. & Bolker, B. M. I can see clearly now: Reinterpreting statistical significance. *Methods Ecol. Evol.* **10**, 756–759 (2019).
4. August, T. *et al.* Emerging technologies for biological recording. *Biol. J. Linn. Soc.* **115**, 731–749 (2015).
5. eBird, T. eBird mobile app for iOS now available! - eBird. https://ebird.org/ebird/news/ebird_mobile_ios1 (2015).
6. eBird, T. Celebrating eBird's 20th Anniversary - eBird. https://ebird.org/ebird/news/ebird-20th-anniversary (2022).
7. 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).
8. 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).
9. Roilo, S., Engler, J. O. & Cord, A. F. Global impact of the COVID-19 lockdown on biodiversity data collection. *Sci. Rep.* **15**, 8767 (2025).
10. Crimmins, T. M., Posthumus, E., Schaffer, S. & Prudic, K. L. COVID-19 impacts on participation in large scale biodiversity-themed community science projects in the United States. *Biol. Conserv.* **256**, 109017 (2021).

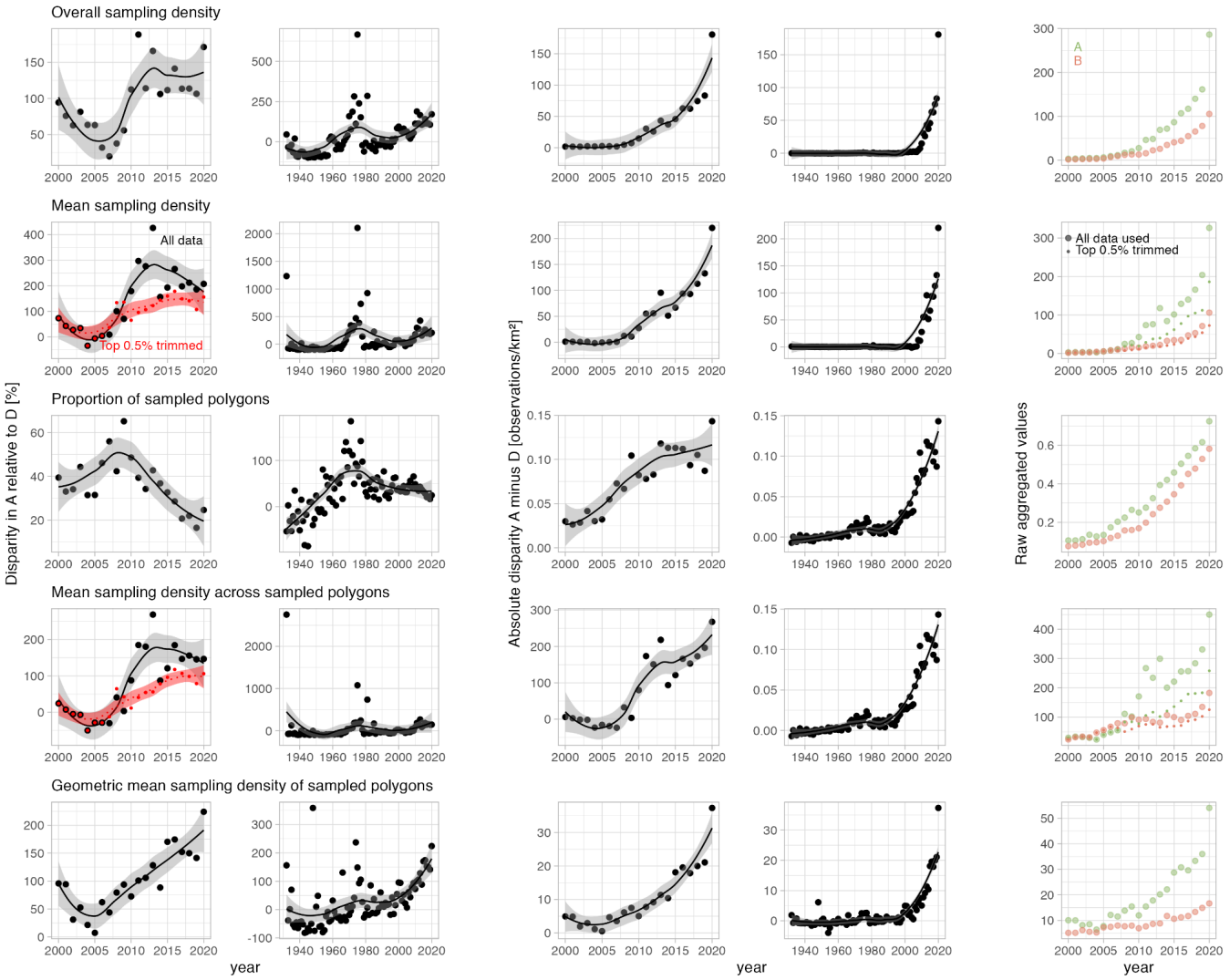


Figure 1 | Change in relative disparity in bird sampling density between HOLC grade A and D over time. Each point represents relative percentage difference (two left columns), absolute difference (two middle columns) in sampling density of A given D (with D being a baseline) based on overall sampling density (i.e. sum of all A or D observations divided by the total area of A or D; **1st row**), mean sampling density per HOLC grade and year (**2nd row**), proportion of sampled neighbourhoods (**3rd row**), mean sampling density across sampled neighbourhoods (i.e. excluding non-sampled ones; **4th row**) and geometric mean in sampling density (**5th 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 in the first four columns represent local regression non-parametric smoothing and shaded areas 95% confidence intervals. Colour in the left column indicates all data (black) or data with top 0.5% trimmed (red), in the right 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 about 35.6% increase, whereas the **other rows** represent alternative aggregations used here. Note that across 2000–2020, alternative metrics (rows 2-5), which represent different sampling mechanisms and down-weight extreme hotspots or isolate distinct sampling processes, reveal a markedly different temporal pattern from the original 35.6% claim (1st row).

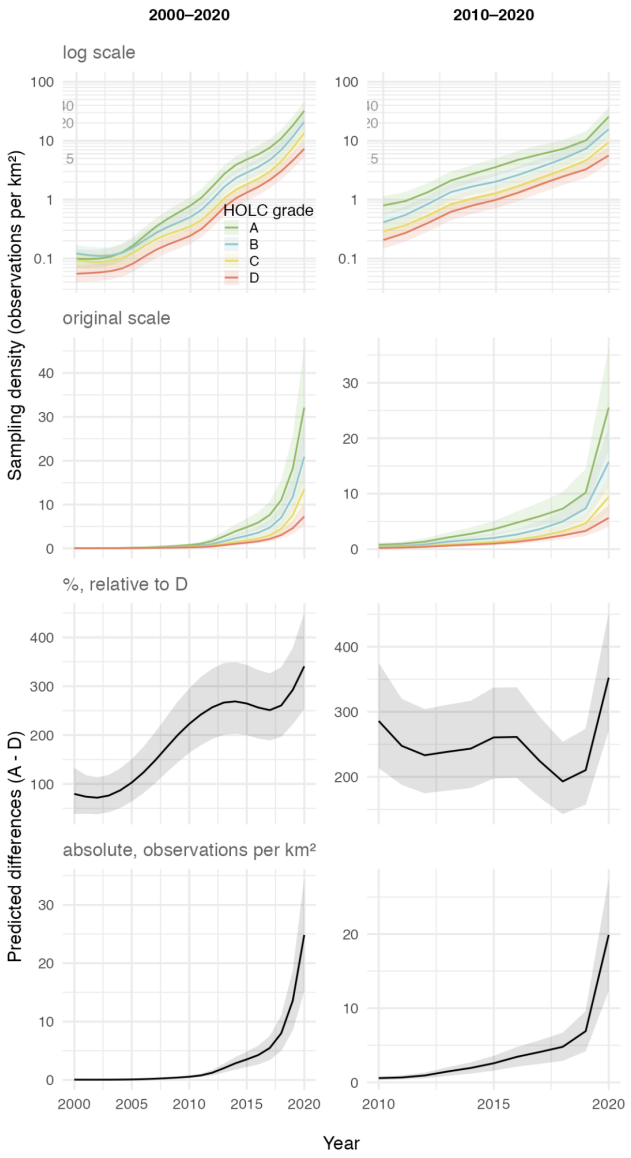
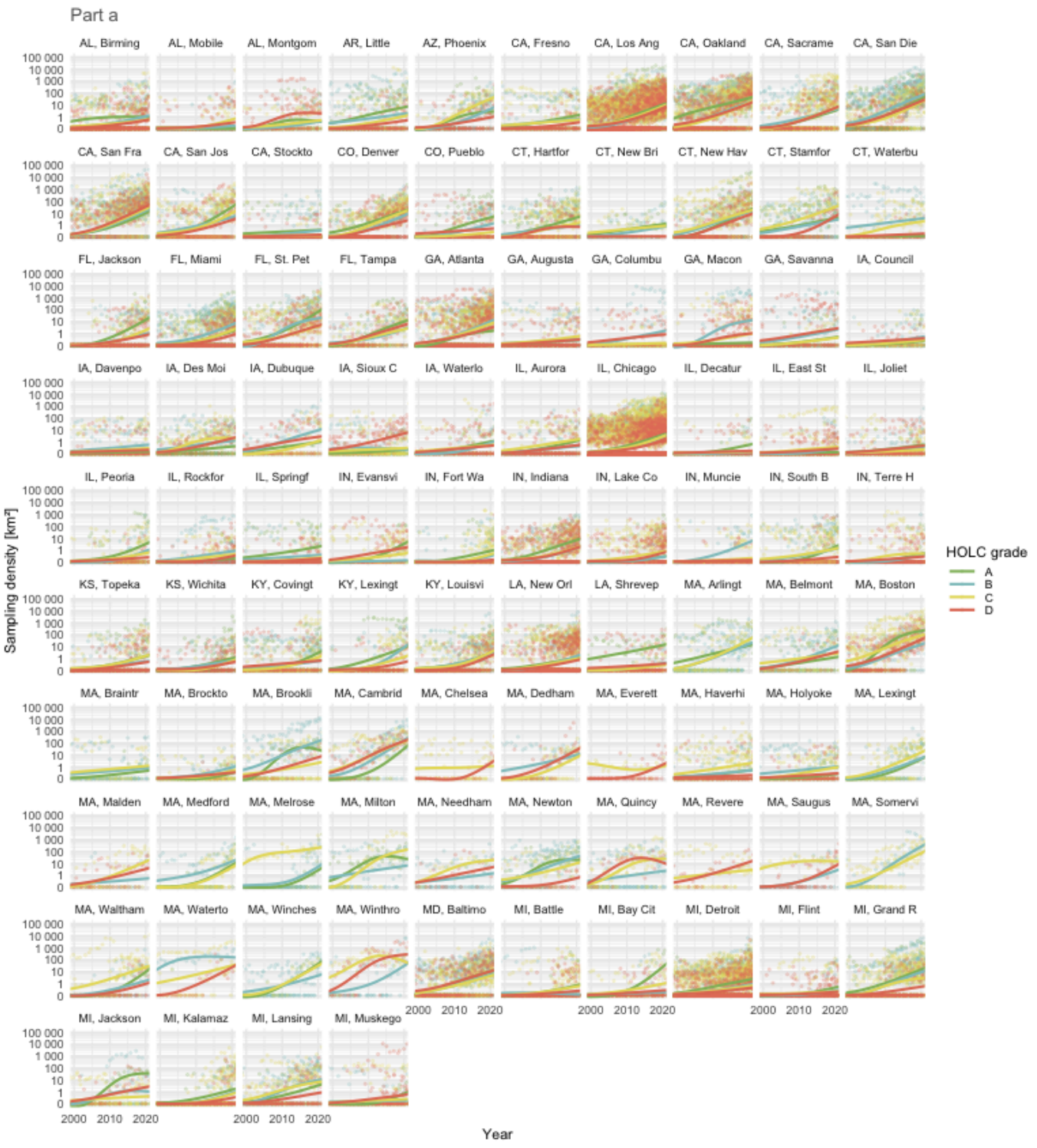
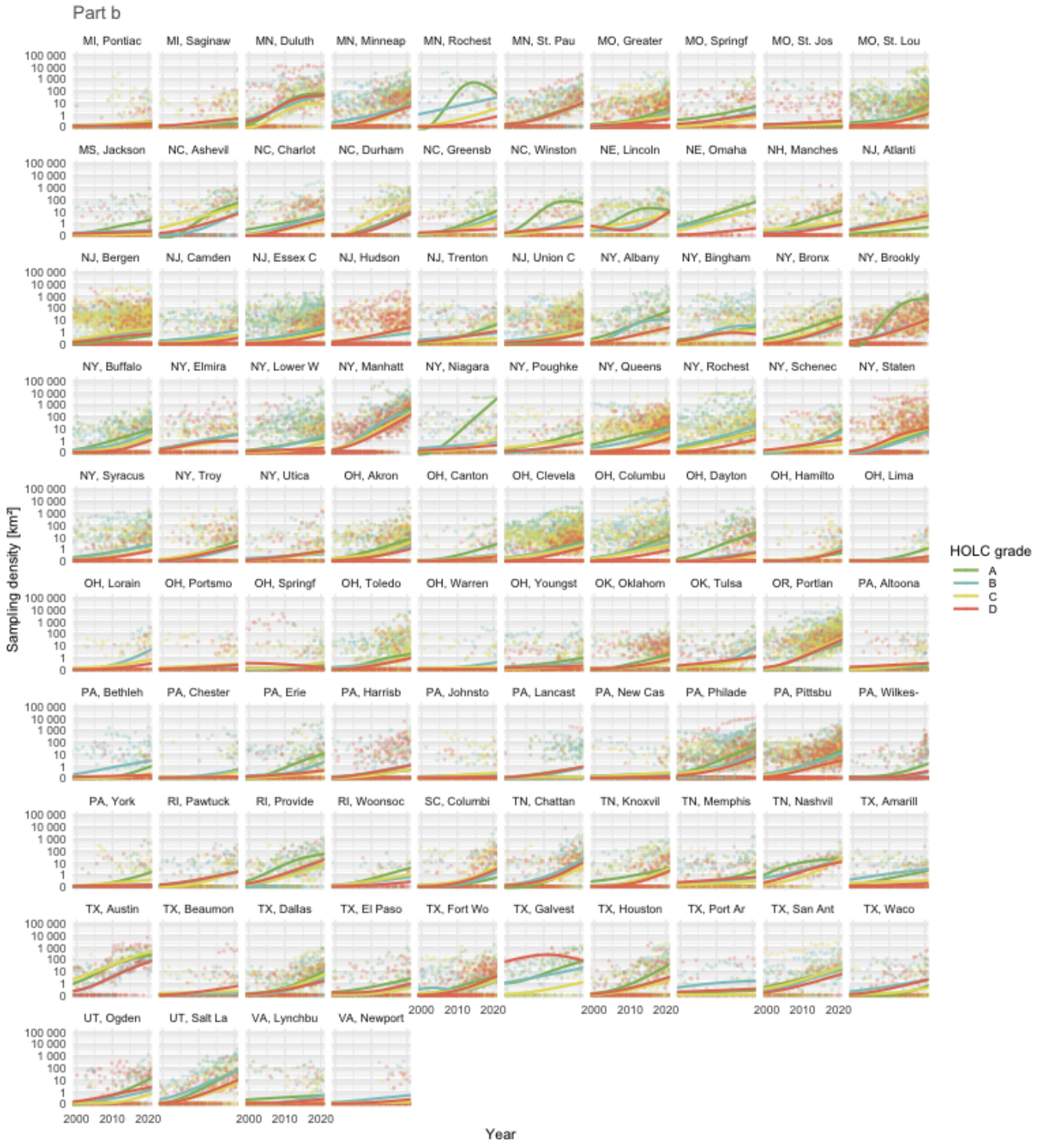
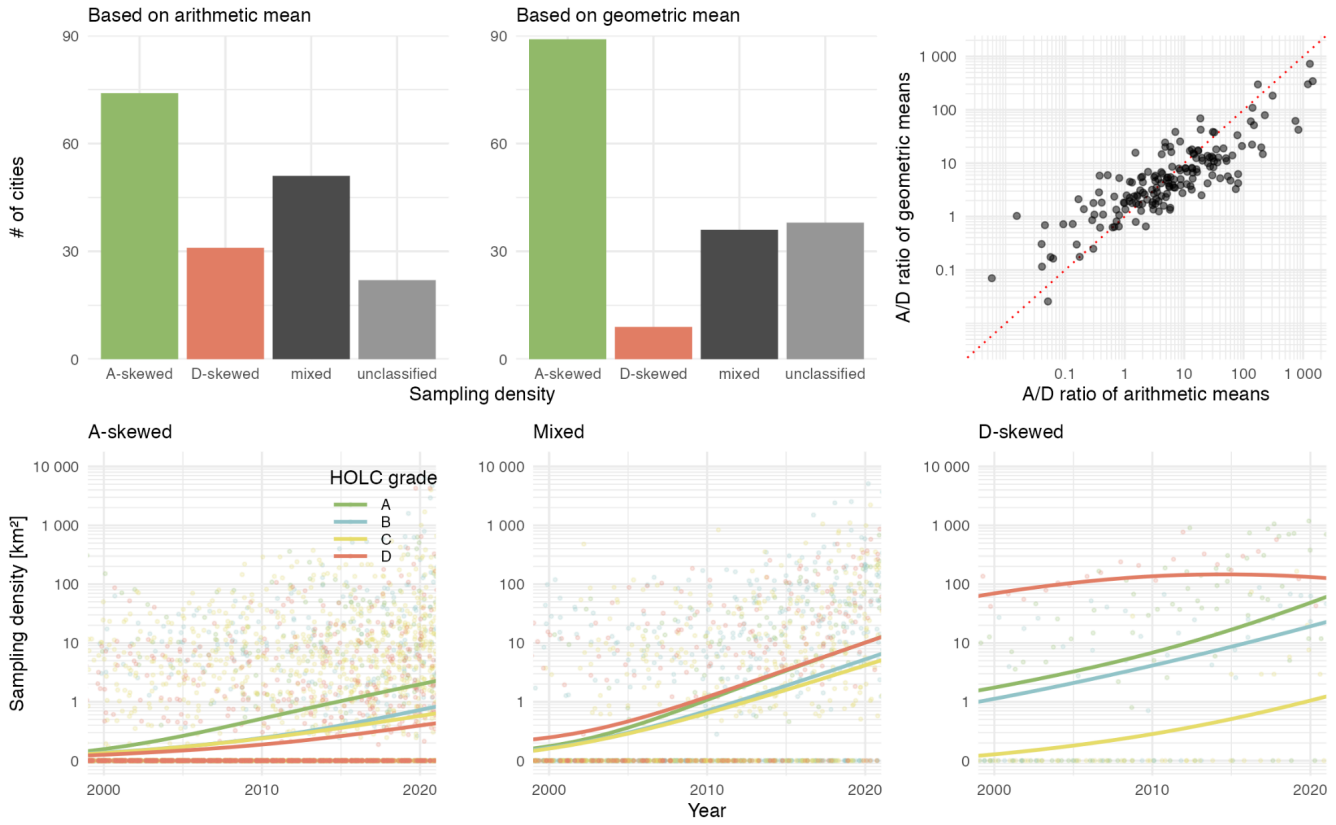


Figure 2 | Non-linear temporal changes in bird-sampling density by HOLC grade and disparity between grades A and D. **Top two rows**, population-level (marginal) predictions from a negative binomial generalised additive model (`bam`) with log link and log(area(km^2)) offset (centered), fitted to neighbourhood-level 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 neighbourhood, 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 reaches ~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.

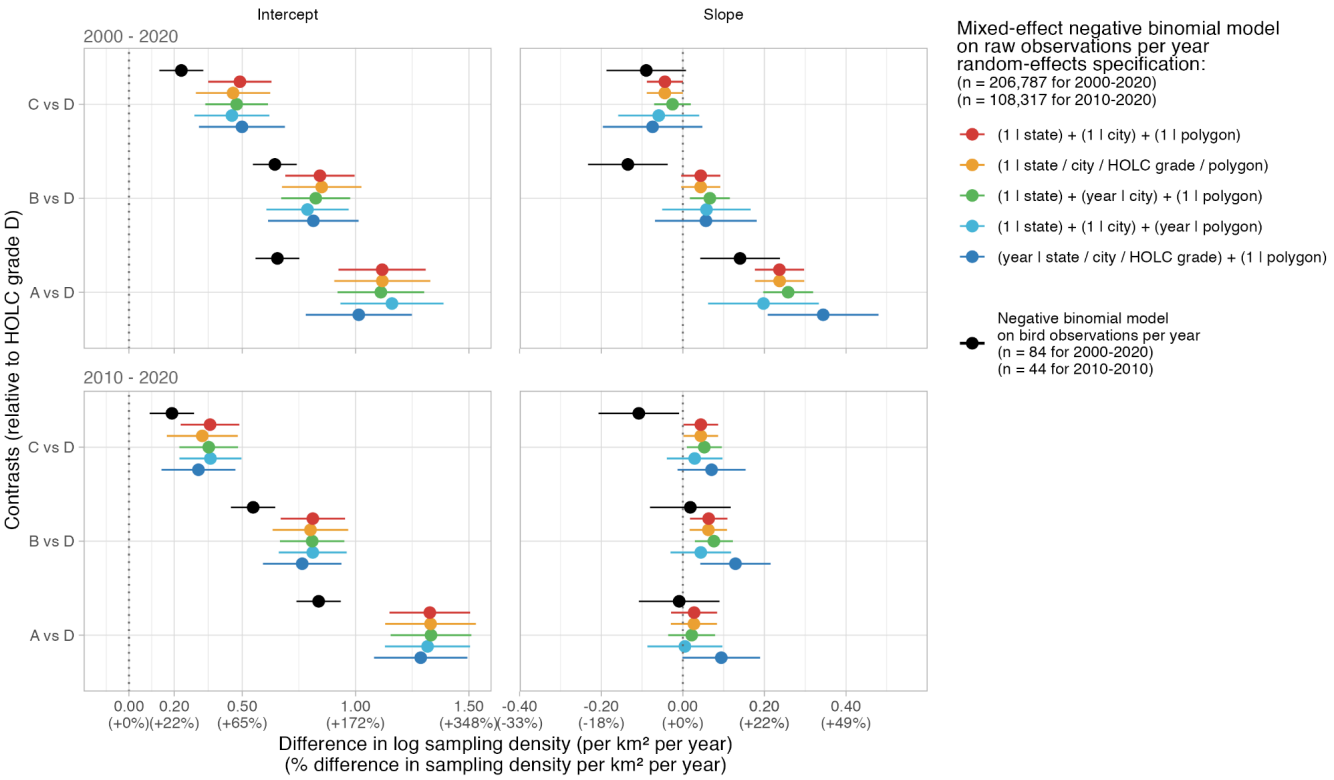

Part a
AL, Birming
AL, Mobile
AL, Montgom
AR, Little
AZ, Phoenix
CA, Fresno
CA, Los Ang
CA, Oakland
CA, Sacrame
CA, San Die
CA, San Fra
CA, San Jos
CA, Stockto
CO, Denver
CO, Pueblo
CT, Hartfor
CT, New Bri
CT, New Hav
CT, Stamfor
CT, Waterbu
FL, Jackson
FL, Miami
FL, St. Pet
FL, Tampa
GA, Atlanta
GA, Augusta
GA, Columbu
GA, Macon
GA, Savanna
IA, Council
IA, Davenpo
IA, Des Moi
IA, Dubuque
IA, Sioux C
IA, Waterlo
IL, Aurora
IL, Chicago
IL, Decatur
IL, East St
IL, Joliet
IL, Peoria
IL, Rockfor
IL, Springf
IN, Evansvi
IN, Fort Wa
IN, Indiana
IN, Lake Co
IN, Muncie
IN, South B
IN, Terre H
KS, Topeka
KS, Wichita
KY, Covingt
KY, Lexingt
KY, Louisvi
LA, New Orl
LA, Shrevep
MA, Arlingt
MA, Belmont
MA, Boston
MA, Braintr
MA, Brockto
MA, Brookli
MA, Cambrid
MA, Chelsea
MA, Dedham
MA, Everett
MA, Haverhi
MA, Holyoke
MA, Lexingt
MA, Malden
MA, Medford
MA, Melrose
MA, Milton
MA, Needham
MA, Newton
MA, Quincy
MA, Revere
MA, Saugus
MA, Somervi
MA, Waltham
MA, Waterto
MA, Winches
MA, Winthro
MD, Baltimo
MI, Battle
MI, Bay Cit
MI, Detroit
MI, Flint
MI, Grand R
MI, Jackson
MI, Kalamaz
MI, Lansing
MI, Muskego
100 000
10 000
1 000
100
10
1
0
2000
2010
2020
Sampling density [km²]
Year
HOLC grade
A
B
C
D



Extended Data Figure 1 | Change in HOLC grade sampling density per km² over time across cities. Lines represent locally estimated scatterplot smoothing or predictions from generalised additive models (generated by stat_smooth function from ggplot2 R-package). Line colour indicates the HOLC grade. Dots depict raw data.



Extended Data Figure 2 | 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 neighbourhoods (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 neighbourhoods ("hotspots"), and therefore reflect occasional survey campaigns more than the underlying spatial structure. Geometric means minimise outliers and capture the "typical" neighbourhood in the "typical" year. The wide scatter around the 1:1 line shows that no single aggregation metric yields a stable classification of cities, 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 neighbourhoods are consistently sampled more densely than D-grade neighbourhoods, mixed city (middle) with no persistent ordering between A and D grades across years, and D-skewed city (right), where D-grade neighbourhoods receive higher sampling density. Panels show raw neighbourhood-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 Fig. 1. The plots for each city are in Extended Data Fig. 1.



Extended Data Figure 3 | 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 km² per year; relative to grade D), obtained by exponentiating those contrasts from negative binomial mixed models with log link and offset of log(area(km²)). Intercept panels show the 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 neighbourhood and year); the **bottom row** contains estimates for a dataset from 2010-2020 (n = 44 and n = 108,317, respectively). Differences between estimates for 2000–2020 and 2010–2020 highlight temporal instability in linear trends, motivating the use of flexible smooth models (Fig. 2).