Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Greenery, sky and water in view and how people judge urban soundscapes

Ruining Zhang
Tianjin University
Hui Ma
Tianjin University
Xinhao Yang
Shenyang Jianzhu University
Chao Wang
Tianjin University
Yuan Zhang
Shenyang Jianzhu University
Jian Kang
j.kang@ucl.ac.uk

University College London

Article

Keywords: Soundscape, Audio-visual interaction, Street-level imagery, Semantic segmentation, Visual quality, Seasonal variation

Posted Date: November 20th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11609344/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: Competing interest reported. Author Jian Kang is Editorial Board Member of Humanities and Social Sciences Communications. Jian Kang was not involved in the journal's review of, or decisions related to, this manuscript. The other authors declare no competing interests.

Greenery, sky and water in view and how people judge urban soundscapes

Ruining Zhang[1,2], Hui Ma[1], Xinhao Yang[2,3], Chao Wang[1], Yuan Zhang[2,3*], Jian Kang[1,4*]

[1]School of Architecture, Tianjin University, Tianjin, 300072, China.
[2]Liaoning Key Laboratory of Eco-Building Physics Technology and Evaluation, Shenyang Jianzhu University, Shenyang, 110168, China.
[3]School of Architecture and Urban Planning, Shenyang Jianzhu University, Shenyang, 110168, China.
[4]UCL Institute for Environmental Design and Engineering, The Bartlett, University College London, London, WC1H 0NN, United Kingdom.

*Corresponding author(s). E-mail(s): y.zhang@sjzu.edu.cn; j.kang@ucl.ac.uk;

Abstract

Street-level imagery is widely used to measure how urban public space looks, and increasingly how it sounds. It measures visual scene composition, namely how much of a view greenery, sky, buildings and other categories occupy. Whether composition captures the visual influences on soundscape evaluation is rarely tested at the same moment. The study analysed 2,262 on-site soundscape questionnaires from 12 urban open public spaces in Shenyang, a northern Chinese city. Each had segmentation and image metrics of the respondent's view and, for 2,229, a binaural recording. Three findings emerged. First, composition explained and predicted almost none of overall evaluation, pleasantness or harmony beyond perceived sound sources and loudness, which explained about thirty times as much variance. This held out of sample and at held-out sites. Eventfulness showed a small predictive gain within sites, associated with visible vehicles, but this gain did not extend to held-out sites. Second, scene condition, the appearance of what is seen, revealed an association that composition missed: greater sky blueness was associated with higher overall evaluation beyond the auditory measures, whereas the amount of sky was not. Greenery share was constant between autumn and spring while its colour content fell by a fifth. Third, visual associations varied with perceived sound characteristics, perceived loudness and the sound sources heard. Visible water helped more as perceived loudness increased, sky blueness

did not, and traffic both heard and seen lowered evaluation beyond its separate effects. Based on these results, scene composition has limited value in explaining or predicting soundscape evaluation. Visual assessment and design should instead focus on sky condition and on specific elements such as water in loud settings.

Keywords: Soundscape, Audio-visual interaction, Street-level imagery, Semantic segmentation, Visual quality, Seasonal variation

1 Introduction

In urban open public space, the visual field is a continuous composition of greenery, sky, ground, buildings, people and traffic. Street-level imagery is now widely used in urban analytics to measure it as such (Biljecki and Ito 2021). Photographs are segmented into scene categories, and the areal shares of these categories form the visual scene composition (Zhou et al. 2019). Composition measures of this kind have been used to predict measured street noise (Zhang et al. 2026), momentary noise perception along daily trajectories (Song et al. 2024), soundscape ratings mapped across whole cities (Zhao et al. 2023; Rui et al. 2024) and wider impressions of place (Le et al. 2025). Estimating how a place sounds from a picture of it depends on an audio-visual premise. The soundscape approach defines the acoustic environment by how it is perceived and understood in context, not by sound level alone (International Organization for Standardization 2014; Cain et al. 2013). It describes soundscape evaluation along two essential dimensions: pleasantness, how agreeable the sound environment is felt to be, and eventfulness, how much is perceived to be happening in it (Axelsson et al. 2010). Context encompasses the non-acoustic components of a soundscape experience (Zhang et al. 2025b). The visible scene represents an observable part of its spatial-physical attributes. A review of soundscape prediction models found no common estimate of the contribution of visual variables (Lionello et al. 2020). What has not been established is how well scene composition explains and predicts soundscape evaluation relative to perceived sound characteristics, the sound sources people hear and how loud they find them. The two have rarely been measured at the same moment for the same person.

The audio-visual premise has wide support. In field surveys and in the laboratory alike, the visible setting is associated with how the acoustic environment is judged (Jeon et al. 2011; Viollon et al. 2002). The association is not uniform: it depends on the type of place and on the sound heard. Conversely, attention recovered differently across park zones that looked broadly alike but were dominated by traffic, mower or natural sound (Zhang et al. 2017). Previous syntheses have situated landscape elements within the spatial-physical context of soundscape experience (Zhang et al. 2025b) and documented associations with pleasantness and eventfulness that vary with site function and element characteristics (Zhang et al. 2025a).

Three limitations of this evidence, however, constrain what it can tell us about scene composition and its implications for assessment and design. It is element-focused: single features are manipulated or noted against controlled or unmeasured backdrops, and in recent field models the visual side is the respondent's own rating of the view

(Lin et al. 2025; Gan et al. 2026). The weight of the whole visible scene relative to the heard one is therefore unknown. It is largely between-place: settings that differ visually are compared, so visual, acoustic and social differences change together and the visual share cannot be isolated. Places that look different also sound different: the diversity of audible sources varied with the panoramic shares of buildings, grass and hard ground (Fu et al. 2024). And it is season-blind: almost all field surveys visit each place once. The state of the scene, the property of urban vegetation and sky that changes most over a year, therefore never varies within the data. The laboratory evidence does not resolve these limits, because it manipulates single elements against controlled backdrops, and its agreement with field responses varies with the reproduction method (Lu and Lau 2025a,b). Neither a soundscape map built from segmented imagery nor a recommendation as basic as "add greenery for a better soundscape" can be evaluated from evidence with these limitations.

Scene composition may also miss the visual features that matter, because areal segmentation is blind to the state of what it segments (Zhou et al. 2019). A bare November canopy and a lush May one return the same greenery share, and a grey and a blue sky the same sky share. We call this state of what is in view the scene condition, as distinct from composition, which records only how much of the view each category covers. Streetscape research has encountered the same problem. The vegetation class includes leafless branches, and the visibly green share of the same streets was far lower in March than in September (Han et al. 2023). The quality or condition of elements matters beside their quantity (Van Renterghem 2019; Zhang et al. 2025b). Attention restoration theory likewise concerns experienced qualities that a visible area does not measure (Kaplan 1995). Image-derived wildness and visual depth were associated with soundscape quality where the variation in green view was not (Zhou et al. 2026). Time spent looking at snow-covered water and leafless vegetation was unrelated to restorative ratings (Zhang et al. 2024). Particular elements, such as water or flowers, raise a related problem: they occupy too little of a view to register as area. If the visual associations with soundscape evaluation depend on scene condition or on such elements rather than on areal quantity, composition metrics will systematically miss them. Both the null results and the positive ones in the audio-visual literature would then need reinterpreting. Testing this requires what field soundscape studies rarely have: repeated surveys of the same sites across seasons, and access to the underlying photographs so that condition can be measured directly rather than inferred.

This study therefore asks how well visual scene composition explains and predicts urban soundscape evaluation, and what matters beyond it. The other visual scene feature considered is scene condition. We address this question through three research questions (RQs):

RQ1. How well does visual scene composition explain and predict different dimensions of soundscape evaluation relative to perceived sound characteristics?

RQ2. Does scene condition reveal associations with soundscape evaluation beyond those captured by scene composition?

RQ3. Do associations between visual scene features and soundscape evaluation vary with perceived sound characteristics?

We answer these questions with a multi-season, photograph-linked soundscape survey of 12 open public spaces in Shenyang, China. Each of 2,262 respondents rated the soundscape they were standing in while the surrounding scene was photographed and, for 2,229 of them, the sound environment was recorded binaurally. Scene composition, scene condition, perceived sound sources and loudness, and measured acoustics therefore describe the same moment at the same place.

2 Methods

The study is a cross-sectional, multi-site field survey of urban open public spaces in Shenyang, a provincial capital in northeastern China with a cold continental climate. Respondents rated the soundscape they were currently experiencing while the surrounding visual scene was photographed and a paired sound recording captured the acoustic environment.

2.1 Sites, campaigns and sample

Data were collected with the Soundscape Indices (SSID) protocol, an in situ questionnaire survey based on Method A of ISO/TS 12913-2:2018 (Mitchell et al. 2020; International Organization for Standardization 2018). The study uses the 12 open public spaces with analysable site photographs: eight parks, three civic squares and one historic workers'-village residential neighbourhood (Table 1). Campaigns ran in eight waves between spring 2019 and spring 2025, covering spring, summer and autumn. Four sites were surveyed in more than one season and provide the seasonal contrasts, two of them over six or more waves. Trained instructors approached visitors at random, obtained informed consent, and each respondent completed a questionnaire, delivered digitally or on paper depending on campaign, taking about ten minutes. Of the 2,792 respondents in the cleaned survey, the 2,262 with an analysable photograph record form the photo-valid analytic sample (50% women; ages 4–93, median 40). One further site, in another city, has no segmented photographs and does not enter this study. All procedures were carried out in accordance with the Declaration of Helsinki and were approved by the Ethics Committee of the School of Architecture and Urban Planning, Shenyang Jianzhu University (No. 20200901). Informed consent was obtained from all participants and, where applicable, their parents or legal guardians.

2.2 Visual-scene measures

Panoramas were photographed from the respondent's position using GoPro Fusion, GoPro Max, Detu F4 or four-sided phone capture, depending on wave. GoPro panoramas were exported as three (Fusion) or four (Max) views in GoPro Player and corrected with Photoshop's Fisheye-Hemi plug-in; other photographs were used without this correction. A PSPNet model with a ResNet-101 backbone trained on ADE20K assigned pixels to 150 scene categories (Zhao et al. 2017; Zhou et al. 2019). Category shares were averaged over each respondent's views. The 2,262 respondents share 2,188 panoramas. Scene composition comprises these areal shares, aggregated into eight

composites: greenery, sky, water, paving, built surface, people, vehicles and street furniture. Together they cover 98% of segmented area; the main exclusions are terrain (1.8%) and animals and residual categories (0.14%). A principal component analysis of the 14 most prevalent categories recovers the greenery-versus-built and openness axes. Raw shares were retained despite 54 profiles summing above 1.05 (maximum 1.26), because renormalised composites correlate with them at $r \geq 0.997$. Sparse elements were represented by presence: water above 0.5% of the view (18% of respondents), flowers and sculpture or fountain each above 0.1% (15% and 11%).

Scene condition describes image-derived appearance. Metrics from 8,414 parseable photographs were averaged over each respondent's views; one respondent lacked matchable images, leaving 2,261. Images were reduced where necessary to 768 pixels on the longer side. In hue–saturation–value space, green pixels have hue 60–170°, saturation ≥ 0.15 and value ≥ 0.08. Their image fraction and its ratio to segmented greenery (defined above 5% greenery) measure vegetation state. Nine condition metrics enter outcome models. Vegetation saturation and hue are the mean saturation and circular mean hue of green pixels, the latter expressed as a shift from 120° towards yellow. Excess green averages $2g - r - b$ on chromaticity coordinates; colourfulness uses the Hasler–Süsstrunk index. Sky blueness is the blue fraction of blue and white pixels in the upper 40% of the image. Blue pixels have hue 185–255°, saturation 0.08–0.75 and value ≥ 0.5; white pixels have saturation below 0.12 and value ≥ 0.75. Brightness and contrast are mean luminance and its standard deviation; edge density is the fraction with luminance gradient above 0.08; luminance entropy uses 64 bins. Green pixels correlate with segmented greenery at $r = 0.80$ (0.71 within sites), and sky pixels with sky share at 0.72 within sites (Supplementary Table S1). Every condition model adjusts for capture system (three-shot, four-shot or dual-generation), which varies within waves. The metrics are comparative rather than colorimetric.

2.3 Constructed visual and audio-visual measures

Visible traffic/road combines road and vehicles; visible activity adds people. Social amenity combines people, seating, tables and playthings. Seven match/conflict indices multiply standardised constituents, controlling their main effects. Three heard-and-seen matches pair traffic salience with visible traffic/road, natural-sound salience with greenery plus water, and human-sound salience with people. Two conflicts pair traffic salience with greenery plus water and natural-sound salience with visible traffic/road. Activity match pairs heard activity (traffic plus human salience) with visible activity; nature-hard balance match pairs auditory with visual nature balance. Auditory balance subtracts mean traffic and other/mechanical salience from natural-sound salience; visual balance subtracts visible traffic/road from greenery plus water. Seven constructed-scene descriptors summarise broader visual context. Diversity and dominance are the Shannon entropy and largest share of the normalised category profile. Natural–hard mixedness measures the balance between greenery, water and terrain versus paving, built surface and vehicles. Open horizon contrasts sky, mountain and water with building, wall, fence and railing shares. Mobility infrastructure combines road and vehicles; social amenity is defined above; restorative cues combine water,

flowers, fountains and benches. The public code specifies the exact formulas, category groupings and image-processing settings.

2.4 Acoustic measurement

While each respondent answered, an instructor wearing binaural microphones stood beside them and recorded the surrounding sound environment. Depending on campaign, the recording system was a B&K 4101A microphone set with a LAN-XI 3160 module, a HEAD acoustics SQuadriga II, or a B&K 2250 / BSWA BR2022 system. Recordings of up to 60 s (30 s from 2024 onwards) were analysed in B&K Connect 23.0.0.935 and HEAD acoustics ArtemiS SUITE 15.1 following ISO/TS 12913-3:2025 (International Organization for Standardization 2025). The extracted indicators were L_{Aeq}, statistical levels, loudness (N, N_5), sharpness, roughness, fluctuation strength and tonality (Fastl and Zwicker 2007). L_{Aeq} is the A-weighted equivalent continuous sound pressure level over the recording ($L_{\mathrm{Aeq,T}}$, with T the recording length) and N_5 the loudness exceeded for 5% of the measurement time. Indicators were computed per channel and enter the analyses as the average of the valid channels, the two-ear average where both channels are valid and the single valid channel otherwise. In the two-channel subsample, the left channel alone correlates with the two-ear average of L_{Aeq} at 0.988, with no mean offset. Acoustic analyses use the 2,229 photo-valid responses with at least one valid channel. The remaining responses enter perception-only models.

2.5 Questionnaire measures

Soundscape evaluation was measured with the eight perceived affective qualities of the Mandarin ISO-12913 instrument (Li et al. 2024). These were projected to pleasantness (*ISO-P*) and eventfulness (*ISO-E*) with the language-specific adjusted angles of the Soundscape Attributes Translation Project and normalised to $[-1, 1]$ (Aletta et al. 2024). The overall sound-environment evaluation (OA), perceived loudness and the perceived audio-visual harmony were each rated on a five-point scale. *ISO-P* and *ISO-E* are the two essential dimensions of the ISO model. OA captures overall appraisal, *ISO-P* captures pleasantness, and harmony describes how well the sound environment fits the visible scene. *ISO-E* describes perceived activity. All four outcomes are analysed throughout. Respondents also rated the perceived salience of traffic noise, human sounds, natural sounds and other/mechanical noise, how noticeable each was, on five-point scales. These four salience ratings and perceived loudness are the five auditory measures through which the perceived sound characteristics of RQ1 and RQ3 enter the models. Four perceived-restorativeness items were averaged into a restorativeness score (RAQ). It enters the match/conflict and constructed-scene families as a fifth outcome, on which no visual or cross-modal term is significant after correction (all $q \geq 0.20$). Person measures are age, gender and education; visitor status (local, tourist, other) was harmonised across the two coding eras of the questionnaire, and visit stage (staying, arriving, leaving) was recorded where collected. Survey wave enters every model as a technical covariate, as wave explains a partial R^2 of 0.08–0.16 in the greenery, sky, built-surface and people composites within sites.

2.6 Statistical analysis

Continuous variables are standardised, giving slopes in SD per SD; element effects are presence contrasts in outcome SD. Linear mixed models have a random site intercept and fixed wave effects, fitted with `lme4` and `lmerTest` in R 4.5 (Bates et al. 2015; Kuznetsova et al. 2017). Each composite, element or metric enters separately. Reported effects are also split into site means and within-site deviations; interactions split both constituents and their product. This matters particularly for vehicles and built surface, whose between-site variance shares are 67% and 58%. Seasonal models use four multi-season sites (autumn 756, spring 851, summer 154), with site and site-by-wave random effects. Perceived-loudness and sound-level moderation use the same structure on the acoustic subsample. Condition models add capture system and, in the full sample, a panorama-group random effect; colour metrics adjust for greenery share and blueness for sky share.

Variance partitions compare the five auditory measures against the eight composites as partial R^2 net of site and wave. ISO coordinates are outcomes, not auditory predictors. Prediction uses grouped 10-fold cross-validation repeated ten times, keeping shared panoramas together to avoid inflating performance (Versümer et al. 2025). Outcomes and predictors are adjusted for site and wave within training folds, using those coefficients for held-out folds. Linear regression and a 500-tree random forest (`ranger`) compare auditory, composition and combined blocks. Forest settings are chosen per outcome and block using the first repeat's grouped cross-validation. Held-out predictions are pooled within each repeat to calculate R^2, then averaged across repeats. Visual gains are tested against permutations of the visual block across panoramas within site-by-wave cells, preserving shared panoramas and using the same folds (1,000 permutations for regression; 200 for forests). Nine condition metrics form a third linear-model block, tested against auditory and combined blocks. Capture system enters both adjustment and permutation cells.

Generalisation to new places is tested by holding out each of the 12 sites. Ordinary least squares models with wave fixed effects use the other 11 sites; all held-out waves occur in training. Models compare auditory, composition, combined and measured-acoustic blocks. The acoustic block has 2,228 complete records for L_{Aeq}, $L_{\mathrm{A10}} - L_{\mathrm{A90}}$, N_5, sharpness, roughness, fluctuation strength and tonality. Person-level R^2 is one minus the ratio of model squared prediction errors to wave-only errors, pooled across held-out sites. Site-level R^2 uses predicted and observed site means. Confidence intervals use 1,000 site-cluster bootstrap resamples. Outcome between-site shares come from site-intercept models adjusted for wave.

Robustness checks use leave-one-site-out refits (12 sites, or four for seasonal contrasts), site-cluster bootstraps (1,000 resamples, dropping failed fits), and element refits restricted to sites where the element varies. Reported effects are also refitted with panorama-group random effects or auditory adjustment. Water models add natural-sound salience or all five auditory measures; water-by-loudness models add the four salience ratings. Exploratory eventfulness pathways screen eight cues: vehicles, visible traffic/road, people, visible activity, social amenity, heard-and-seen traffic, traffic-in-nature conflict and nature-hard balance match. Cue and indirect-path models use

ordinary least squares with site and wave fixed effects; indirect effects are coefficient products. Site-cluster bootstrap intervals are estimated for three cues (vehicles, heard-and-seen traffic and nature-hard balance match) through eventfulness to OA.

Coefficient tests were two-sided; permutation tests for predictive gains used the upper tail of the null distribution. Benjamini–Hochberg correction applies separately within 19 pre-specified families (Benjamini and Hochberg 1995), with significance at $q < 0.05$. The largest families are composition (40 tests), condition metrics (36), match/conflict and constructed scene (35 each), and curvature (32). Next are greenery structure (24), weather (20), source by season (12), congruence (9), and elements and eventfulness cues (8 each). The smallest are perceived-loudness moderation, L_{Aeq} moderation, visitor familiarity and visit stage (7 each), chromatic season contrasts (6), greenery by season (5), blueness by season (3), and blueness by loudness (2). Congruence tests the three heard-and-seen interactions on OA, *ISO-P* and harmony. Their re-standardised forms also enter the match/conflict family, corrected separately. The composition family includes perceived loudness as a fifth outcome; match/conflict and constructed-scene families include RAQ. Nominal results ($p < 0.05$, $q \geq 0.05$) support no claim. Full computational settings are in the public code. Figures use `ggplot2`.

During the preparation of this work the authors used the latest Fable and Opus models on Claude Code and the latest Astra and Sol models on Codex for language and grammar polish, R and Python coding for data analysis and visualisation, and LaTeX formatting. After using these tools, the authors reviewed and edited the content as needed and take full responsibility for the content of the publication.

3 Results

3.1 Scene composition barely explains or predicts evaluation, except on eventfulness

RQ1 asked how well visual scene composition explains and predicts different dimensions of soundscape evaluation relative to perceived sound characteristics. Figure 1 gives the answer: for overall evaluation, pleasantness and harmony, almost not at all. In the common in-sample frame, the auditory measures explain 14–20% of variance in these three outcomes. The eight composites explain 0.4–0.7%, a visual-to-auditory ratio of 0.02–0.04 (Figure 1a). This is not an artefact of the estimator or the functional form. The eight composites form the visual block and the five auditory measures the auditory block. Out of sample, adding the visual block to the auditory one does not improve prediction on any of the three outcomes (gains in cross-validated R^2 of -0.004 to -0.001, permutation $p = 0.10$–0.56; Figure 1b; Supplementary Tables S2 and S3). The visual block alone does not predict these outcomes out of sample (cross-validated $R^2 \leq 0.003$), so the null result is not produced by conditioning on what is heard. A random forest, which predicts better than the linear model from the auditory measures, gains nothing from the visual block either (gains of -0.002 to $+0.0001$, permutation $p = 0.16$–0.33). No composite shows a non-linear dose–response (32 quadratic terms, all $q \geq 0.70$; Supplementary Table S4).

Composition does describe what kind of place one is in. Site-mean greenery correlates with the distinction between parks and other spaces at $r = 0.87$. Three composites (greenery, built surface and sky) classify a respondent's space as park or not at an in-sample area under the curve of 0.85. Its association with evaluation is weak even across places. The greenery slope on *ISO-P* is +0.05 ($p = 0.046$) in a person-level model adjusted only for wave and +0.04 ($p = 0.19$) once a site intercept removes the between-place contrast.

Each site was held out in turn and predicted by models fitted to the other 11. The composites predict nothing about the respondents of the held-out site beyond what the survey wave predicts (held-out R^2 of −0.008 to 0.000; Figure 1c; Supplementary Table S5). The auditory measures reach 0.13–0.19 on the three outcomes, and the measured acoustic indicators are close to zero (−0.004 to 0.005). Adding the composites to the auditory measures improves held-out prediction detectably only for harmony, the rating of how well the sound suits the scene (+0.013, 95% CI [0.004, 0.031]). This gain of about one percentage point is not distinguishable from zero without the wave terms (Supplementary Table S5). Little of the variance in evaluation lies between places, 2.5–3.4% for each outcome, and neither block predicts the site means of overall evaluation, pleasantness or harmony reliably.

Eventfulness is the exception. Its visual-to-auditory ratio is 0.71, although both blocks are small on this outcome (1.6% and 2.2% of variance; Figure 1a). The visual increment remains out of sample in the linear model (+0.007 cross-validated R^2, permutation $p = 0.001$; Figure 1b). The forest's gain is smaller and not significant (+0.002, $p = 0.070$). At held-out sites the increment disappears (−0.010, 95% CI [−0.030, 0.007]; Supplementary Table S5), so this visual signal holds within places rather than across them. Of the 40 tests in the composition family, only visible vehicles on *ISO-E* remains significant (+0.102, $q < 0.001$; Figure 2a, Table 2; the full family is Supplementary Table S6). This effect is cross-modal and persists essentially unchanged under adjustment for L_{Aeq} (+0.090, $p < 0.001$) and for the perceived sound-source salience ratings. Audio-visual match and conflict indices, which combine scene composition with perceived sound sources while controlling their constituent main effects, act on the same dimension (Figure 2b; Supplementary Table S7). Traffic heard together with visible traffic and the alignment of the auditory with the visual nature–hardscape balance both raise *ISO-E* ($q \leq 0.018$), the latter robust under site-cluster bootstrap. Traffic sound inside a visually natural frame lowers it. Eight visual and cross-modal cues were taken forward as candidate routes from the scene through *ISO-E* into overall evaluation (Supplementary Table S8). None is a useful route: their indirect paths range from −0.007 to +0.005 (Supplementary Table S9). Of the three paths bootstrapped, the largest, for visible vehicles, spans −0.013 to +0.001, and the only interval that excludes zero, for the balance match, is negative. Eventfulness therefore reflects both the seen and the heard scene, acting as an endpoint rather than a mediator into overall evaluation.

The same holds for richer descriptions of scene quantity. The internal structure of greenery (tree, grass and herbaceous fractions, flower content, layering, diversity) adds nothing on overall evaluation, pleasantness or harmony beyond amount (best $q = 0.35$; Supplementary Table S10). A seven-parameter block of constructed scene descriptors

(openness, diversity, mixedness, infrastructure, affordance and related composites) has no significant effect after correction across 35 tests (best $q = 0.203$; Supplementary Table S11). It adds 0.2–0.5% of variance beyond the eight composites.

3.2 Sky condition is associated with evaluation and greenery condition changes with season

RQ2 asked whether scene condition reveals associations with soundscape evaluation beyond those captured by scene composition. Three sets of results address this question (Figure 3).

First, the segmentation layer is season-blind while the scene itself is not. Across the four multi-season sites, the greenery share is essentially identical in autumn and spring (0.28 versus 0.31; Figure 3a): a bare canopy segments like a lush one. The photographs show what the segmentation misses: green pixels per unit of segmented greenery fall from 0.53 in spring to 0.42 in autumn (Figure 3c). The contrast holds under site-cluster bootstrap (+0.11, 95% CI [0.07, 0.23]). The seasonal difference in vegetation colour is therefore measured rather than inferred.

Second, the greenery slope on overall evaluation appears in spring, although its seasonal difference is not established. The within-site greenery slope on overall evaluation is +0.13 in spring ($p = 0.004$) and essentially zero in autumn (−0.003; Figure 3b; season-stratified estimates in Supplementary Table S12). The greenery-by-season interaction is nominal only ($p = 0.023$, $q = 0.113$; Supplementary Table S13) and supports no claim by itself. *ISO-P* shows no seasonal difference (+0.11 in spring against +0.08 in autumn). The seasonal pattern is clear on the auditory side (Figure 3d). The benefit of perceived natural sound on overall evaluation is +0.25 in spring and +0.20 in summer against +0.05 in autumn (interaction $q = 0.018$). The interaction is also significant on harmony and *ISO-P* (both $q = 0.0497$; Supplementary Tables S14 and S15), while the negative effect of traffic noise shows no significant seasonal difference (joint $q = 0.157$). The natural-sound contrast on overall evaluation is unchanged when each of the four sites is left out in turn. The seasonal pattern also holds under adjustment for daily temperature and air quality (joint $p = 0.002$–0.018 across the natural-sound interactions, with the spring contrast on overall evaluation at $p = 3.8\times10^{-4}$, and joint $p = 0.020$ for greenery; Supplementary Table S16). Date-level weather and air-quality main effects are themselves null (all $q \geq 0.81$; Supplementary Table S17), so daily temperature and air quality do not explain the seasonal contrasts. Chromatic vegetation metrics do not, however, explain the spring slope. The seasonal colour contrasts are individually indistinguishable from zero once equipment is adjusted for ($q = 0.24$–0.32; Supplementary Table S18). No vegetation-colour metric remains significant in the condition-metric family (best $q = 0.124$ across the four colour metrics; Supplementary Table S19), and the greenery-by-season interaction is not attenuated when colour enters the model. The mechanism behind the seasonal contrast therefore remains open even though the vegetation condition itself is now measured.

Third, a pure condition variable is associated with evaluation where the matching quantity is not. Of the 36 tests in the condition-metric family, only sky blueness on overall evaluation remains significant (+0.103, $q = 0.006$; Table 2; the full family is

Supplementary Table S19). The sky *share* has no effect on overall evaluation, pleasantness or harmony that remains significant after correction (on overall evaluation $p = 0.032$, $q = 0.21$). Once blueness is in the model it has none ($p = 0.58$). The blueness effect is robust. It is sign-stable under leave-one-site-out refits, has a site-cluster bootstrap 95% CI of [0.03, 0.18], and is unchanged by blur and clipping filters (+0.112) and by daily-weather adjustment (+0.111). It persists at +0.066 ($p = 0.007$) with all the auditory measures in the model (Supplementary Table S20), so it is not a proxy for what is heard. Its seasonal profile matches the greenery pattern, with the slope concentrated in spring (+0.119, $p = 0.003$) rather than autumn (+0.023), although this contrast is nominal ($q = 0.111$). Autumn skies are no less blue, so this nominal difference is in the slope and not in the amount of blue sky. As a set, however, the nine condition metrics improve out-of-sample prediction no more than composition does. Added to the auditory measures, their gain in cross-validated R^2 lies between -0.006 and -0.0004 (permutation $p \geq 0.08$). Added to the auditory and composition blocks together, it is at most +0.0005, a borderline gain on *ISO-E* (permutation $p = 0.04995$; Supplementary Tables S2 and S3). The condition effect is therefore specific to sky blueness and is not a general property of the photograph metrics.

3.3 Visual associations vary selectively with perceived sound characteristics

RQ3 asked whether associations between visual scene features and soundscape evaluation vary with perceived sound characteristics. Beyond the eight areal composites and sky blueness, the elements recorded as present or absent (water, flowers, sculpture or fountain) were first tested in the eight-test element family. We then examined how the visual associations vary with the two perceived sound characteristics, perceived loudness and which sound sources are heard. Water in view is the clearest case (Figure 4a). Its presence effect on overall evaluation is +0.20 ($q = 0.005$), with +0.17 on *ISO-P* (Table 2). It is the only season-robust element effect, positive in every season stratum (+0.27 in autumn, $p = 0.005$; +0.20 in spring; +0.44 in the small summer stratum, $p = 0.032$). Its association increases with perceived loudness (interaction +0.167, $q = 0.010$; Supplementary Table S21). Stratifying the acoustic sample into L_{Aeq} tertiles gives a water slope of +0.42 in the loud tertile ($p < 0.001$). In the quiet and mid tertiles the slopes are +0.08 and +0.15, neither distinguishable from zero (Supplementary Table S22). The corresponding interaction with metered L_{Aeq} is not significant (+0.09, $p = 0.109$; Supplementary Table S23), so the compensation finding depends on the perceived-loudness interaction. The water benefit is largest where the sound environment is experienced as loudest, a pattern consistent with visual compensation.

The sky-blueness association provides the contrast case (Figure 4b). Its interaction with perceived loudness is indistinguishable from zero ($q = 0.76$). Greater sky blueness is associated with higher evaluation across the loudness range rather than specifically under high perceived loudness, a profile distinct from water's. Flowers, the remaining significant element in the presence family (+0.23 on overall evaluation, $q = 0.005$; Table 2), also do not vary with perceived loudness and are strongly seasonal. The slope reaches +0.44 ($p = 0.003$) in spring, when flowering is meaningful, and is

indistinguishable from zero in autumn. As the effect is concentrated in one season, flowers are a supporting rather than a principal result (Supplementary Table S24).

Variation with perceived sound sources is likewise source-specific rather than general (Figure 4c). Traffic heard together with visible traffic lowers overall evaluation beyond the two main effects (-0.057, $q = 0.036$, and $q = 0.035$ within the larger match/conflict family). The effect is concentrated in autumn (-0.12, $p = 0.005$, against -0.02 in spring) and is sign-stable under leave-one-site-out refits. No corresponding general congruence benefit appears for nature or people, so cross-modal combination matters for overall evaluation where the same aversive source is both heard and seen, not as a general matching principle.

Four checks qualify these effects. The strongest effects are spatially concentrated, with the vehicle–eventfulness signal largest at the residential neighbourhood site and the water effect at the one major site whose water is widely visible. Under site-restricted refits, which re-estimate each element only at the sites where it varies, all effects keep their direction. Only the flower effect, however, keeps a site-bootstrap interval excluding zero (95% CI [0.16, 0.36]; Supplementary Table S25). The water and vehicle results are therefore within-site effects whose generality across settings needs testing at more sites. Seven visual slopes, including the vehicle, water and flower effects, were also tested for differences between respondent groups. None varies with visitor familiarity (best $q = 0.35$ across the harmonised local/tourist contrast) or with visit stage (staying versus passing through, all $q \geq 0.67$; Supplementary Tables S26 and S27). Adding a panorama-group random effect changes each estimate by less than 0.004. When each effect is split into a between-site and a within-site component, every within-site component is significant ($p \leq 0.020$) and close to the reported estimate (Supplementary Table S28). Finally, the water results were refitted with what is heard in the model (Table 2). The interaction with perceived loudness changes little when the four sound-source salience ratings are added ($+0.157$, $p = 0.002$). The average water effect is partly shared with the auditory measures. With natural-sound salience in the model, under which water sound is rated, the effect on overall evaluation falls to $+0.16$ ($p = 0.007$). With all five auditory measures it falls to $+0.10$ ($p = 0.074$).

4 Discussion

4.1 Principal findings and comparison with previous work

Scene composition, the areal profile that segmentation-based urban analytics measures, explained and predicted almost none of overall evaluation, pleasantness or harmony once the heard environment was accounted for. A composition model fitted at some places also predicted nothing at others. Places that look different also sound different (Fu et al. 2024), so an association between view and evaluation across places can reflect the acoustic and social differences that accompany the view. Here the places themselves differed little in evaluation, and most of its variance lay between people and moments at the same place. What a person hears varies within a place in a way that a picture of the place does not record. This is why the perceived sources still predicted evaluation at places the models had not seen. Within a place, at a matched moment and with the heard environment in the model, the areal shares had

little left to explain. Composition describes the kind of place a person is in, not how its soundscape is judged.

The present results reframe the positive audio-visual findings rather than contradict them. In laboratory pairings of sound and view, changing the sound affected the judgement more than changing the view (Preis et al. 2015; Van Renterghem and Lippens 2024). Perceived sound sources have also predicted pleasantness much better than mapped spatial context, although the two were never combined in one model (Barros et al. 2026). The pooled association of vegetation with pleasantness is strong but highly heterogeneous (Zhang et al. 2025a), and several visual associations changed once the acoustic environment was adjusted for (Zhou et al. 2026). Those positive results are based largely on element presence, manipulated stimuli or between-place contrasts. Because composition is exactly what street-view segmentation pipelines measure at scale (Zhao et al. 2023; Rui et al. 2024), such pipelines can be expected to predict soundscape evaluation mainly through what the view indicates about the sound. The indication is imperfect, since the sources judged to suit a silent scene agree only moderately with those later heard in it (Jo and Jeon 2020).

Eventfulness was the one dimension that scene composition predicted, through visible vehicles, while its routes into overall evaluation were negligible. A plausible reason is that the two ISO dimensions have different sensory bases. Eventfulness describes how much is happening in a place, and visible movement, vehicles above all, is a sign of activity in its own right. Pleasantness is a judgement of how good the sound is, and at a fixed place that judgement follows what is heard. In the individual encounter with a place, then, the sense of activity is multisensory while valence is dominated by hearing. In judgements made from silent street images, visible scene features predicted the eventfulness that respondents expected (Zhao et al. 2023). Crowds and the sky are the elements associated with eventfulness, and vegetation with pleasantness (Zhang et al. 2025a). Eventfulness also has an auditory basis of its own, in the dominance of human sounds (Axelsson et al. 2010) and in how sound events co-occur (Karges et al. 2026). For imagery-based soundscape sensing the implication is an asymmetry. A composition model has a small signal of its own for how eventful a place feels. For how pleasant a place sounds, it has only a proxy signal, through the co-variation of what is seen with what is heard.

Scene condition was associated with evaluation in one respect that the data establish: greater sky blueness was associated with higher evaluation, whereas the amount of sky was not. Earlier studies measured the sky as an amount, and the reviewed findings range from positive to null (Zhang et al. 2025b). The seasonal results are compatible with a wider role of condition but do not establish it. Greenery was valued in spring and natural sound in the growing seasons, but the vegetation colour metrics did not themselves predict evaluation. A plausible explanation is a seasonal appraisal state in which natural content of either modality is valued when the environment signals growth. Natural sound and natural scenery can reinforce each other, although virtual-reality experiments differ on whether their combination exceeds the two separate effects (Kawai et al. 2025; Jin et al. 2026). That natural sound was valued in the seasons when the vegetation was in leaf suggests a response shared across the two senses. The state of planting also matters to its appraisal: flower cover was the

strongest correlate of the aesthetic response to designed planting (Hoyle et al. 2017), which agrees with the flower effect found here. The value of natural content in a scene may thus depend on its state as well as on its area. Streetscape research measures greenery as share and treats its seasonal swing as a measurement problem (Han et al. 2023). For perception research, the swing is itself a variable worth measuring.

Associations between visual scene features and soundscape evaluation depended on perceived sound characteristics. Visible water improved evaluation most where perceived loudness was highest, whereas the sky-blueness benefit did not depend on perceived loudness, and traffic lowered evaluation further only when it was both heard and seen. A visual element can thus act on soundscape evaluation as an amenity, valued across loudness levels, or as a remedy, valued most where the sound environment is experienced as loudest. Visible water rarely comes without its sound. In the classic stress-recovery experiment the water scene was shown with its stream sound (Ulrich et al. 1991). In a virtual-reality study, added water sounds were judged most appropriate at the one site with water in view (Hong et al. 2020). Part of the average benefit of visible water here was indeed shared with how the place sounded, but its dependence on perceived loudness was not. Visual context may matter more at higher exposure, although the evidence is mixed (Van Renterghem 2019). The opposite pattern has also been reported: the association of visual nature with lower negative affect disappeared at the highest level of anthropogenic sound (Aldoh et al. 2025). The value of what is seen thus depends on the perceived sound environment in which it is seen. This gives the preference and restorative value attached to scenes containing water (White et al. 2010; Jeon et al. 2023) a specific acoustic reading.

The heard-and-seen traffic result qualifies the idea of audio-visual congruence in the same way. Seeing the road made traffic noise more appropriate in the virtual-reality study, with little gain in pleasantness (Hong et al. 2020). Here cross-modal combination mattered where it combined two doses of the same aversive source, and no general benefit of audio-visual matching appeared. Congruence is therefore a property of particular source combinations, not a general design principle.

4.2 Implications for assessment and design

For assessment, the results caution against relying on segmented scene composition to explain or predict how a soundscape is evaluated. The panoramas here were taken where each respondent stood and reduced to the same kind of areal shares that street-view pipelines compute. The caution therefore concerns the measure itself rather than a particular image source. A soundscape map built on such shares predicts differences between places. Where evaluation varies mainly between people and moments at the same place, those differences are small, and imagery-based maps are better suited to describing places than to estimating how their soundscapes are judged. A monitoring pipeline built on areal shares will report a leafless park and a lush one as visually identical, and will miss the sky condition that was associated with evaluation here. Static imagery also cannot show transient, hidden or distant sources (Chen et al. 2026). Where imagery is used to audit open public space for perceptual outcomes, sky state is a cheap supplement that these data support. Colour content per unit vegetation records a seasonal change that areal shares cannot detect, although the vegetation

colour metrics did not predict evaluation here. Measures of where elements are in the view (Feng and Li 2026) or of how far the view extends (Llorca-Bofí et al. 2024) are further ways beyond areal composition. Two provisos apply to any condition metric. It inherits camera and processing differences, as image-derived measures generally do (Biljecki and Ito 2021), so a repeated audit needs an equipment term or a fixed capture standard. It is also seasonal by nature, so a single-season audit samples a state, not a property. Eventfulness deserves separate handling: it is the only outcome for which scene composition had predictive value of its own, and that value was small and did not extend to places outside the training data.

For design, the measures this study supports concern specific elements and their placement, not the area of each scene category. Adding greenery area to a space that already looks green is unlikely to change how its soundscape is judged. The seasonal results indicate when natural content is valued, in spring and summer, and so inform the planning of use across seasons. They do not identify a vegetation measure that design can target, because no vegetation-condition metric predicted evaluation. Water has a specific role: placed where perceived loudness is highest, its visible presence may complement the use of water sound over traffic noise. The response to water sound depends on its type (Ji et al. 2025) and on whether the site is also seen (Hong et al. 2020). Where traffic cannot be quieted, screening the sight of the source would remove the combined heard-and-seen exposure. Sky access, finally, is not a matter of quantity. Openness programmes justified on soundscape grounds find no support here, because the sky effect depended on the condition of the sky and not on its amount.

4.3 Limitations

Six limitations qualify these conclusions. First, the design is observational and cross-sectional: within-site identification and matched-moment measurement exclude many confounds, but no visual factor was experimentally assigned. Second, the auditory measures and the outcomes are ratings by the same respondent at the same moment, whereas the visual measures come from photographs. Part of the auditory share therefore reflects the common instrument, and the thirty-fold ratio compares two sets of measures, not the contributions of the two senses. Third, the multi-season contrasts are based on four sites, and season is partially entangled with survey batch. The leave-one-site-out refits and the weather adjustment support the natural-sound contrast, but a season-balanced replication would settle it. Fourth, the water and vehicle effects each concentrate at one site, so their generality across settings is not established, and composites identify within-site effects only. Fifth, the photograph metrics are comparative rather than colorimetric. Camera and processing generations shift them, the capture-system proxy corrects for the shift only to first order, and the chromatic mechanism behind the spring slope remains unconfirmed. Sixth, the sample is 12 open public spaces in one northern Chinese city and one survey language. Mean ratings have not been shown to be equivalent across translations of the instrument (Aletta et al. 2024), and the translated and an indigenous Chinese scale gave differently distributed scores (Wu et al. 2025). Absolute levels should therefore not be exported.

5 Conclusions

Across 2,262 photograph-linked soundscape assessments in 12 urban open public spaces in Shenyang, China, this study tested how well visual scene composition explains and predicts urban soundscape evaluation. Areal scene composition has limited value in explaining or predicting overall evaluation, pleasantness and harmony, not because vision is irrelevant, but because the visual associations that remain come from scene condition and specific elements and vary with perceived sound characteristics.

On RQ1, visual scene composition explains and predicts almost none of overall evaluation, pleasantness or harmony beyond perceived sound characteristics, which explain around thirty times as much variance. The result holds out of sample and in prediction at held-out sites. For eventfulness, the visual and auditory contributions are comparable and both small; its visual predictive gain is confined to within-site validation and does not extend to held-out sites. Visible vehicles have the only significant compositional association with it ($q < 0.001$), and audio-visual match indices also predict it without providing an indirect pathway into overall evaluation.

On RQ2, scene condition reveals an association with evaluation that scene composition misses: sky blueness raises overall evaluation, robustly and beyond all the auditory measures, whereas sky amount does not. Segmented greenery share is constant between autumn and spring while its measured colour content drops by a fifth. The benefit of perceived natural sound is specific to spring and summer and the greenery slope on overall evaluation appears in spring, while the negative effect of traffic noise shows no significant seasonal difference (joint $q = 0.157$). The nine condition metrics as a set improve out-of-sample prediction no more than composition does, so the condition effect established here is specific to sky blueness.

On RQ3, associations between visual scene features and soundscape evaluation vary selectively with perceived sound characteristics. The benefit of visible water increases with perceived loudness and is largest in the loudest tertile, consistent with visual compensation. The sky-blueness benefit does not vary with loudness, and the one negative cross-modal interaction on overall evaluation occurs when traffic is both heard and seen.

Segmented scene composition should therefore not be relied on alone to explain or predict soundscape evaluation, and image-based assessment of open public space should look beyond composition wherever perceptual outcomes are the target. Urban visual design can support soundscapes through specific elements and where they are placed, more than through how much of each scene category the view contains. Prospective designs that assign scene condition experimentally, segmentation models that output state alongside class, and season-balanced multi-city replication are the natural extensions of this evidence.

Data availability. The original data presented in the study are openly available in Mendeley Data at https://doi.org/10.17632/wr5hrbpjbs.

Code availability. The analysis code that reproduces the figures and tables is openly available on GitHub at https://github.com/SJZU-Acoustics/scene-composition-soundscape.

Acknowledgements. This study was supported by the Educational Department of Liaoning Province Project (No. JYTZD2023172) and the European Research Council (ERC) Advanced Grant (No. 740696).

Author contributions. **Ruining Zhang:** Writing – Original Draft, Methodology (lead), Investigation (lead), Formal analysis (lead). **Hui Ma:** Writing – Review & Editing (supporting), Supervision (equal), Methodology (supporting). **Xinhao Yang:** Investigation (supporting). **Chao Wang:** Formal analysis (supporting). **Yuan Zhang:** Writing – Review & Editing (lead), Supervision (equal), Methodology (supporting), Funding acquisition (equal), Conceptualization (equal). **Jian Kang:** Writing – Review & Editing (supporting), Supervision (equal), Methodology (supporting), Funding acquisition (equal), Conceptualization (equal).

Competing interests. Author Jian Kang is Editorial Board Member of *Humanities and Social Sciences Communications*. Jian Kang was not involved in the journal's review of, or decisions related to, this manuscript. The other authors declare no competing interests.

Ethics declarations. All procedures were carried out in accordance with the Declaration of Helsinki and were approved by the Ethics Committee of the School of Architecture and Urban Planning, Shenyang Jianzhu University (No. 20200901). Informed consent was obtained from all participants and, where applicable, their parents or legal guardians.

References

Aldoh A, Ungureanu R, Popescu S, et al (2025) How multisensory nature exposure affects well-being: Visual and auditory influences on affect. People and Nature 7(11):2821–2837. https://doi.org/10.1002/pan3.70157

Aletta F, Mitchell A, Oberman T, et al (2024) Soundscape descriptors in eighteen languages: Translation and validation through listening experiments. Applied Acoustics 224:110109. https://doi.org/10.1016/j.apacoust.2024.110109

Axelsson Ö, Nilsson ME, Berglund B (2010) A principal components model of soundscape perception. The Journal of the Acoustical Society of America 128(5):2836–2846. https://doi.org/10.1121/1.3493436

Barros A, Van Renterghem T, Lembrechts JJ, et al (2026) From place to perception: the influence of geospatial context on soundscape assessment. Building and Environment 295:114445. https://doi.org/10.1016/j.buildenv.2026.114445

Bates D, Mächler M, Bolker BM, et al (2015) Fitting linear mixed-effects models using lme4. Journal of Statistical Software 67(1):1–48. https://doi.org/10.18637/jss.v067.i01

Benjamini Y, Hochberg Y (1995) Controlling the false discovery rate: A practical and powerful approach to multiple testing. Journal of the Royal Statistical Society: Series B (Methodological) 57(1):289–300. https://doi.org/10.1111/j.2517-6161.1995.tb02031.x

Biljecki F, Ito K (2021) Street view imagery in urban analytics and GIS: A review. Landscape and Urban Planning 215:104217. https://doi.org/10.1016/j.landurbplan.2021.104217

Cain R, Jennings P, Poxon J (2013) The development and application of the emotional dimensions of a soundscape. Applied Acoustics 74(2):232–239. https://doi.org/10.1016/j.apacoust.2011.11.006

Chen P, Huang X, Fei T, et al (2026) Cross-modal urban sensing: Evaluating sound–vision alignment across street-level and aerial imagery. Transactions in GIS 30(2):e70246. https://doi.org/10.1111/tgis.70246

Fastl H, Zwicker E (2007) Psychoacoustics: Facts and Models, 3rd edn. Springer, Berlin, Heidelberg, https://doi.org/10.1007/978-3-540-68888-4

Feng S, Li R (2026) Dynamic audio-visual preference in waterfront walking environments: Effects of distance-layered spatial structure and soundscape. Building and Environment 304:115060. https://doi.org/10.1016/j.buildenv.2026.115060

Fu W, Huang J, Ran C, et al (2024) How the landscape pattern and landscape elements affect the source-based soundscape diversity. Urban Ecosystems 27(6):2469–2486. https://doi.org/10.1007/s11252-024-01590-6

Gan Y, Li J, Huang J, et al (2026) Context dependent sensory integration modulates landscape preference in urban parks. Scientific Reports 16(1):21270. https://doi.org/10.1038/s41598-026-52401-1

Han Y, Zhong T, Yeh AGO, et al (2023) Mapping seasonal changes of street greenery using multi-temporal street-view images. Sustainable Cities and Society 92:104498. https://doi.org/10.1016/j.scs.2023.104498

Hong JY, Lam B, Ong ZT, et al (2020) Effects of contexts in urban residential areas on the pleasantness and appropriateness of natural sounds. Sustainable Cities and Society 63:102475. https://doi.org/10.1016/j.scs.2020.102475

Hoyle H, Hitchmough J, Jorgensen A (2017) All about the 'wow factor'? The relationships between aesthetics, restorative effect and perceived biodiversity in designed urban planting. Landscape and Urban Planning 164:109–123. https://doi.org/10.1016/j.landurbplan.2017.03.011

International Organization for Standardization (2014) ISO 12913-1:2014 Acoustics — Soundscape — Part 1: Definition and conceptual framework. International

Standard, Geneva, Switzerland

International Organization for Standardization (2018) ISO/TS 12913-2:2018 Acoustics — Soundscape — Part 2: Data collection and reporting requirements. Technical Specification, Geneva, Switzerland

International Organization for Standardization (2025) ISO/TS 12913-3:2025 Acoustics — Soundscape — Part 3: Data analysis. Technical Specification, Geneva, Switzerland

Jeon JY, Lee PJ, Hong JY, et al (2011) Non-auditory factors affecting urban soundscape evaluation. The Journal of the Acoustical Society of America 130(6):3761–3770. https://doi.org/10.1121/1.3652902

Jeon JY, Jo HI, Lee K (2023) Psycho-physiological restoration with audio-visual interactions through virtual reality simulations of soundscape and landscape experiences in urban, waterfront, and green environments. Sustainable Cities and Society 99:104929. https://doi.org/10.1016/j.scs.2023.104929

Ji R, Li S, Bai Z, et al (2025) Are natural soundscapes always beneficial? Evaluating the restorative qualities and influencing mechanisms of natural water soundscapes. Applied Acoustics 227:110205. https://doi.org/10.1016/j.apacoust.2024.110205

Jin T, Zhang Z, Shao Y (2026) Audiovisual modulation of traffic noise effects on psychological restoration in expressway-adjacent residential environments: A virtual reality study. Buildings 16(4):873. https://doi.org/10.3390/buildings16040873

Jo HI, Jeon JY (2020) Effect of the appropriateness of sound environment on urban soundscape assessment. Building and Environment 179:106975. https://doi.org/10.1016/j.buildenv.2020.106975

Kaplan S (1995) The restorative benefits of nature: Toward an integrative framework. Journal of Environmental Psychology 15(3):169–182. https://doi.org/10.1016/0272-4944(95)90001-2

Karges N, Haselhoff T, Taubenböck H, et al (2026) Co-occurrence network analysis of urban acoustic environments: Structural archetypes and perceived quality. Computers, Environment and Urban Systems 127:102437. https://doi.org/10.1016/j.compenvurbsys.2026.102437

Kawai C, Georgiou F, Pieren R, et al (2025) In the ear of the beholder: Restorative potential of different audio-visual characteristics of recreational spaces tested in 360° VR. Journal of Environmental Psychology 107:102760. https://doi.org/10.1016/j.jenvp.2025.102760

Kuznetsova A, Brockhoff PB, Christensen RHB (2017) lmerTest package: Tests in linear mixed effects models. Journal of Statistical Software 82(13):1–26. https://doi.

org/10.18637/jss.v082.i13

Le QH, Dinh NNH, Kim B, et al (2025) Scalable transfer learning framework for capturing human perceptions of place through visual-aural data integration. Cities 166:106286. https://doi.org/10.1016/j.cities.2025.106286

Li M, Han R, Xie H, et al (2024) Mandarin Chinese translation of the ISO-12913 soundscape attributes to investigate the mechanism of soundscape perception in urban open spaces. Applied Acoustics 215:109728. https://doi.org/10.1016/j.apacoust.2023.109728

Lin M, Chau CK, Tang SK, et al (2025) Determinants of soundscape quality of communal open space in Hong Kong. Building and Environment 267:112261. https://doi.org/10.1016/j.buildenv.2024.112261

Lionello M, Aletta F, Kang J (2020) A systematic review of prediction models for the experience of urban soundscapes. Applied Acoustics 170:107479. https://doi.org/10.1016/j.apacoust.2020.107479

Llorca-Bofí J, Heck J, Vorländer M (2024) Relating 2D isovists to audiovisual assessments of two urban spaces in a neighbourhood. Noise Mapping 11(1):20240013. https://doi.org/10.1515/noise-2024-0013

Lu Y, Lau SK (2025a) The ecological validity of laboratory experiments in soundscape and landscape research: A systematic review and meta-analysis. Applied Acoustics 232:110582. https://doi.org/10.1016/j.apacoust.2025.110582

Lu Y, Lau SK (2025b) Examining the ecological validity of VR experiments in soundscape and landscape research. Computers in Human Behavior 162:108462. https://doi.org/10.1016/j.chb.2024.108462

Mitchell A, Oberman T, Aletta F, et al (2020) The soundscape indices (SSID) protocol: A method for urban soundscape surveys—questionnaires with acoustical and contextual information. Applied Sciences 10(7):2397. https://doi.org/10.3390/app10072397

Preis A, Kociński J, Hafke-Dys H, et al (2015) Audio-visual interactions in environment assessment. Science of the Total Environment 523:191–200. https://doi.org/10.1016/j.scitotenv.2015.03.128

Rui Q, Gu K, Cheng H (2024) Integrating street-view images to quantify the urban soundscape: Case study of Fuzhou City's main urban area. The Journal of the Acoustical Society of America 156(4):2090–2105. https://doi.org/10.1121/10.0029026

Song L, Liu D, Kwan MP, et al (2024) Machine-based understanding of noise perception in urban environments using mobility-based sensing data. Computers, Environment and Urban Systems 114:102204. https://doi.org/10.1016/j.compenvurbsys.2024.102204

Ulrich RS, Simons RF, Losito BD, et al (1991) Stress recovery during exposure to natural and urban environments. Journal of Environmental Psychology 11(3):201–230. https://doi.org/10.1016/S0272-4944(05)80184-7

Van Renterghem T (2019) Towards explaining the positive effect of vegetation on the perception of environmental noise. Urban Forestry & Urban Greening 40:133–144. https://doi.org/10.1016/j.ufug.2018.03.007

Van Renterghem T, Lippens A (2024) The audio-visual incongruency asymmetry. Natural sounds in an urban visual setting are more relaxing than urban sounds in visual nature. Urban Forestry & Urban Greening 101:128514. https://doi.org/10.1016/j.ufug.2024.128514

Versümer S, Blättermann P, Rosenthal F, et al (2025) A comparison of methods for modeling soundscape dimensions based on different datasets. The Journal of the Acoustical Society of America 157(1):234–255. https://doi.org/10.1121/10.0034849

Viollon S, Lavandier C, Drake C (2002) Influence of visual setting on sound ratings in an urban environment. Applied Acoustics 63(5):493–511. https://doi.org/10.1016/S0003-682X(01)00053-6

White M, Smith A, Humphryes K, et al (2010) Blue space: The importance of water for preference, affect, and restorativeness ratings of natural and built scenes. Journal of Environmental Psychology 30(4):482–493. https://doi.org/10.1016/j.jenvp.2010.04.004

Wu D, Han R, Zhang R, et al (2025) The untranslatability of environmental affective scales: insights from indigenous soundscape perceptions in China. npj Urban Sustainability 5(1):38. https://doi.org/10.1038/s42949-025-00228-6

Zhang R, Zhu L, Yang X, et al (2024) The effects of interaction with audiovisual elements on perceived restoration in urban parks in freezing weather. Urban Forestry & Urban Greening 96:128337. https://doi.org/10.1016/j.ufug.2024.128337

Zhang R, Ma H, Wang C, et al (2025a) The associations between landscape elements and soundscape perception: A meta-analysis. Landscape and Urban Planning 263:105463. https://doi.org/10.1016/j.landurbplan.2025.105463

Zhang R, Ma H, Wang C, et al (2025b) Soundscape and its context: A framework based on a systematic review. The Journal of the Acoustical Society of America 157(6):4417–4436. https://doi.org/10.1121/10.0036882

Zhang Y, Kang J, Kang J (2017) Effects of soundscape on the environmental restoration in urban natural environments. Noise & Health 19(87):65–72. https://doi.org/10.4103/nah.NAH_73_16

Zhang Y, Ke E, Kwan MP, et al (2026) Multi-frequency street-level urban noise modeling and mapping through street view and remote sensing image fusion. Computers, Environment and Urban Systems 126:102401. https://doi.org/10.1016/j.compenvurbsys.2026.102401

Zhao H, Shi J, Qi X, et al (2017) Pyramid scene parsing network. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp 6230–6239, https://doi.org/10.1109/CVPR.2017.660

Zhao T, Liang X, Tu W, et al (2023) Sensing urban soundscapes from street view imagery. Computers, Environment and Urban Systems 99:101915. https://doi.org/10.1016/j.compenvurbsys.2022.101915

Zhou B, Zhao H, Puig X, et al (2019) Semantic understanding of scenes through the ADE20K dataset. International Journal of Computer Vision 127(3):302–321. https://doi.org/10.1007/s11263-018-1140-0

Zhou Z, Tan PY, Lau SK, et al (2026) Exploring relationships between qualitative attributes of urban greenery and soundscape quality assessment: A case study in Singapore. Journal of Environmental Management 410:130034. https://doi.org/10.1016/j.jenvman.2026.130034

Figure legends. Figure 1. Scene composition barely explains or predicts overall evaluation, pleasantness or harmony. (a) Variance in each evaluation outcome explained by the auditory measures versus the eight scene composites, as partial R^2 net of site and wave in a common person-level frame. The auditory measures are the four perceived sound-source salience ratings and perceived loudness; the sample is 2,262 photo-valid respondents; OA is overall evaluation, *ISO-P* pleasantness and *ISO-E* eventfulness. (b) Out-of-sample gain in cross-validated R^2 from adding the visual block (the eight composites) to the auditory block (the five auditory measures), from grouped 10-fold cross-validation repeated ten times with folds respecting shared panoramas. Orange bands span the 5th–95th percentiles of each outcome's 1,000-permutation null (visual block permuted). The filled point marks the one increment with permutation $p < 0.05$ (*ISO-E*, +0.007, $p = 0.001$). (c) Prediction at held-out sites: person-level R^2 of the auditory measures and the composites (colours as in a), measured against the wave-only model. Each of the 12 sites is held out in turn and predicted by linear models with wave fixed effects fitted to the other 11. Points are estimates and bars 95% CIs from 1,000 site-cluster bootstrap resamples.

Figure 2. Eventfulness is the one evaluation dimension that scene composition predicts, alone and combined with perceived sound sources. (a) Slopes of the eight scene composites on eventfulness (*ISO-E*) from wave-adjusted site-random-intercept linear mixed models ($n = 2{,}262$; points are estimates, bars 95% CIs).

The filled point remains significant after Benjamini–Hochberg correction in the 40-test composition family ($q < 0.001$). (b) Slopes of the seven audio-visual match/conflict indices on *ISO-E*, each controlling its constituent main effects (same sample and model frame; filled points remain significant within the 35-test family, $q < 0.05$).

Figure 3. Greenery amount is constant between seasons while its condition and its association with evaluation change. All panels use the four multi-season sites (autumn 756, spring 851, summer 154 respondents), and the models in (b) and (d) have site and site-by-wave random effects. (a) Mean greenery areal share by season; the summer stratum is small and site-skewed. (b) Within-site greenery slope on overall evaluation (OA) by season (points are estimates, bars 95% CIs; filled points $p < 0.05$). (c) Green pixels per unit of segmented greenery by season, over the respondents whose greenery share exceeds 5% ($n = 751$ autumn, 756 spring photometric respondents). The autumn–spring contrast, from a model with a site random intercept and the capture-system proxy, holds under the site-cluster bootstrap (+0.11, 95% CI [0.07, 0.23]). (d) Within-site slopes of perceived natural sound and traffic noise on OA by season. The natural-by-season interaction remains significant within its 12-test family ($q = 0.018$), and the traffic slope shows no significant seasonal difference (joint $q = 0.157$).

Figure 4. Visual associations across perceived sound characteristics. (a) Water-in-view effect on overall evaluation (OA) within L_{Aeq} tertiles of the photo-valid acoustic sample ($n = 743$ per tertile; points are presence contrasts in outcome SD, bars 95% CIs, filled $p < 0.05$). The water-by-perceived-loudness interaction is +0.167 ($q = 0.010$ in the 7-test moderation family). (b) The sky-blueness effect on OA across specifications. The first is the condition-metric family estimate ($n = 2{,}261$; $q = 0.006$ in the 36-test family), and the second adds all auditory measures to the model (sound-adjusted). The others are the spring and autumn subsets and the interaction with perceived loudness ($n = 2{,}228$). All models adjust for sky share and the capture-system proxy. The first two also include wave fixed effects and site and panorama-group random intercepts. The season subsets have site and site-by-wave random intercepts, and the loudness model adds wave fixed effects to these. (c) The heard-and-seen traffic interaction on OA, pooled ($n = 2{,}262$; $q = 0.036$ in the 9-test congruence family) and within the autumn and spring strata of the multi-season subset.

Table 1 The 12 sites: space class, survey waves, sample sizes, mean scene composition and acoustic level. n is the number of photo-valid analytic respondents and n_{ac} the subsample with at least one valid recording channel. Greenery and Sky are site-mean areal shares from semantic segmentation. OA is the site-mean overall sound-environment evaluation on the raw 1–5 scale. L_{Aeq} is the site-mean A-weighted equivalent level in dB(A) over respondents with valid recordings. Wave labels abbreviate season (Sp spring, Su summer, Au autumn).

Site	Space class	Survey waves	n	n_{ac}	Greenery	Sky	OA	L_{Aeq}
Zhongshan Park	Park	2019 Sp, 2019 Su, 2020 Au, 2021 Au, 2023 Sp, 2024 Sp, 2025 Sp	586	553	0.43	0.12	3.45	64.0
Changbaidao Park	Park	2024 Sp	113	113	0.47	0.12	3.68	60.0
Quan Park	Park	2023 Sp	86	86	0.47	0.17	3.55	62.3
Nanhu Park	Park	2025 Sp	54	54	0.45	0.10	3.69	59.2
Laomo Park	Park	2024 Sp	53	53	0.48	0.08	3.19	58.9
Linli Park	Park	2024 Sp	51	51	0.33	0.17	3.63	62.8
Luoshiquan Ecological Park	Park	2024 Sp	49	49	0.40	0.19	3.86	60.8
Beiling Park	Park	2025 Sp	48	48	0.45	0.18	3.75	54.9
Olympic Park	Civic square	2019 Sp, 2019 Su, 2020 Au, 2021 Au, 2024 Sp, 2025 Sp	507	507	0.30	0.21	3.61	60.2
Zhongshan Square	Civic square	2019 Sp, 2019 Au, 2020 Au, 2024 Sp	345	345	0.12	0.26	3.27	61.0
Dadong Square	Civic square	2019 Sp, 2019 Au, 2020 Au, 2024 Sp	323	323	0.29	0.16	3.41	63.0
Gongren Xincun	Residential	2024 Sp	47	47	0.17	0.09	3.79	59.0
Total			2,262	2,229				

Table 2 Effects that remain significant after Benjamini–Hochberg correction within their pre-specified families. Estimates are in SD per SD for standardised predictors and in outcome SD for presence contrasts. Every model includes wave fixed effects and a site random intercept, except the loudness-moderation model, which includes site and site-by-wave random effects on the acoustic subsample. The condition-metric model additionally includes the sky share, the capture-system proxy and a panorama-group random effect, and moderation models the constituent main effects. Family sizes: composition 40, condition metrics 36, element 8, loudness moderation 7, congruence 9, match/conflict 35. Indented rows are robustness refits of the row above and have no separate q: L_{Aeq}-adjusted adds metered level, sound-adjusted adds all five auditory measures, and source-adjusted adds the four perceived sound-source salience ratings. n = 2,262 (photo-valid), 2,261 (condition metrics) or 2,229 (L_{Aeq}-adjusted and loudness moderation). Two further match/conflict effects on *ISO-E*, heard-and-seen traffic (+0.070) and traffic-in-nature conflict (−0.072), also remain significant (both q = 0.014; Supplementary Table S7).

Effect	Estimate	SE	p	q
Scene composition (per SD area share)				
Visible vehicles → *ISO-E*	0.102	0.024	<0.001	<0.001
the same, L_{Aeq}-adjusted	0.090	0.024	<0.001	–
Scene condition (per SD condition metric)				
Sky blueness → OA	0.103	0.027	<0.001	0.006
the same, sound-adjusted	0.066	0.024	0.007	–
Visual elements (present vs absent)				
Water in view → OA	0.198	0.061	0.001	0.005
the same, sound-adjusted	0.099	0.055	0.074	–
Water in view → *ISO-P*	0.174	0.060	0.004	0.009
Flowers in view → OA	0.228	0.069	<0.001	0.005
Perceived sound characteristics: perceived loudness				
Water × perceived loudness → OA	0.167	0.052	0.001	0.010
the same, source-adjusted	0.157	0.050	0.002	–
Perceived sound characteristics: sound-source combination				
Heard-and-seen traffic → OA	−0.057	0.020	0.004	0.036
Nature-hard balance match → *ISO-E*	0.067	0.021	0.001	0.017

Figures

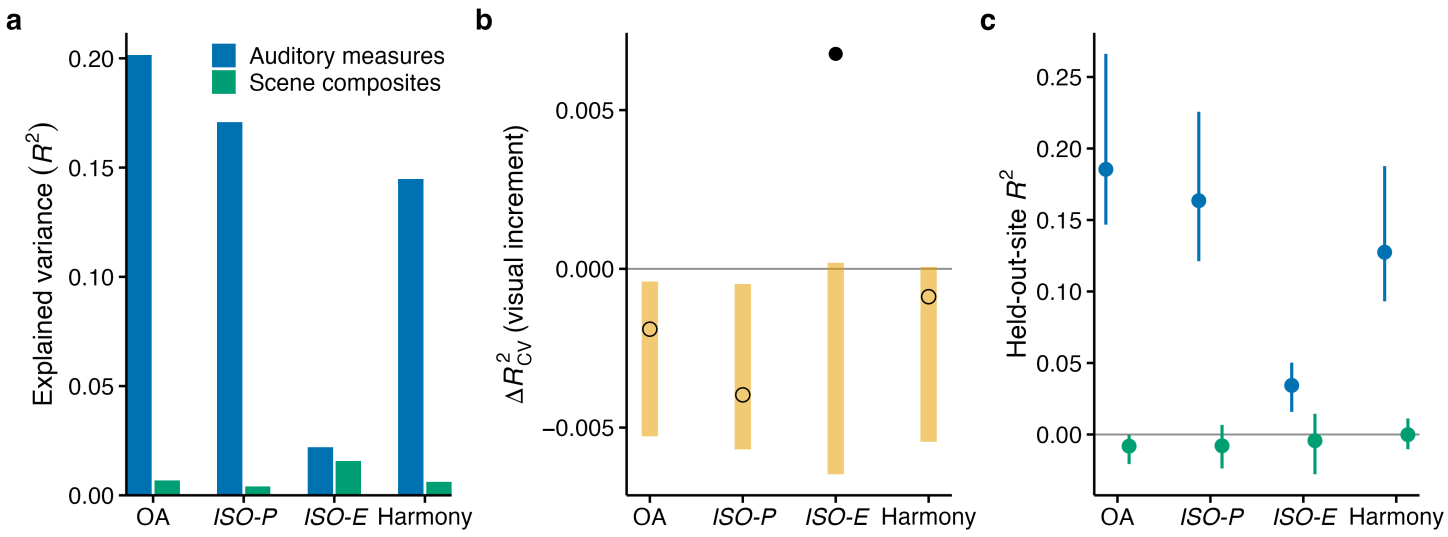


Figure 1

Scene composition barely explains or predicts overall evaluation, pleasantness or harmony. (a) Variance in each evaluation outcome explained by the auditory measures versus the eight scene composites, as partial R2 net of site and wave in a common person-level frame. The auditory measures are the four perceived sound-source salience ratings and perceived loudness; the sample is 2,262 photo-valid respondents; OA is overall evaluation, ISO-P pleasantness and ISO-E eventfulness. (b) Out-of-sample gain in cross-validated R2 from adding the visual block (the eight composites) to the auditory block (the five auditory measures), from grouped 10-fold cross-validation repeated ten times with folds respecting shared panoramas. Orange bands span the 5th–95th percentiles of each outcome's 1,000- permutation null (visual block permuted). The filled point marks the one increment with permutation $p < 0.05$ (ISO-E, +0.007, $p = 0.001$). (c) Prediction at held-out sites: person-level R2 of the auditory measures and the composites (colours as in a), measured against the wave-only model. Each of the 12 sites is held out in turn and predicted by linear models with wave fixed effects fitted to the other 11. Points are estimates and bars 95% CIs from 1,000 site-cluster bootstrap resamples.

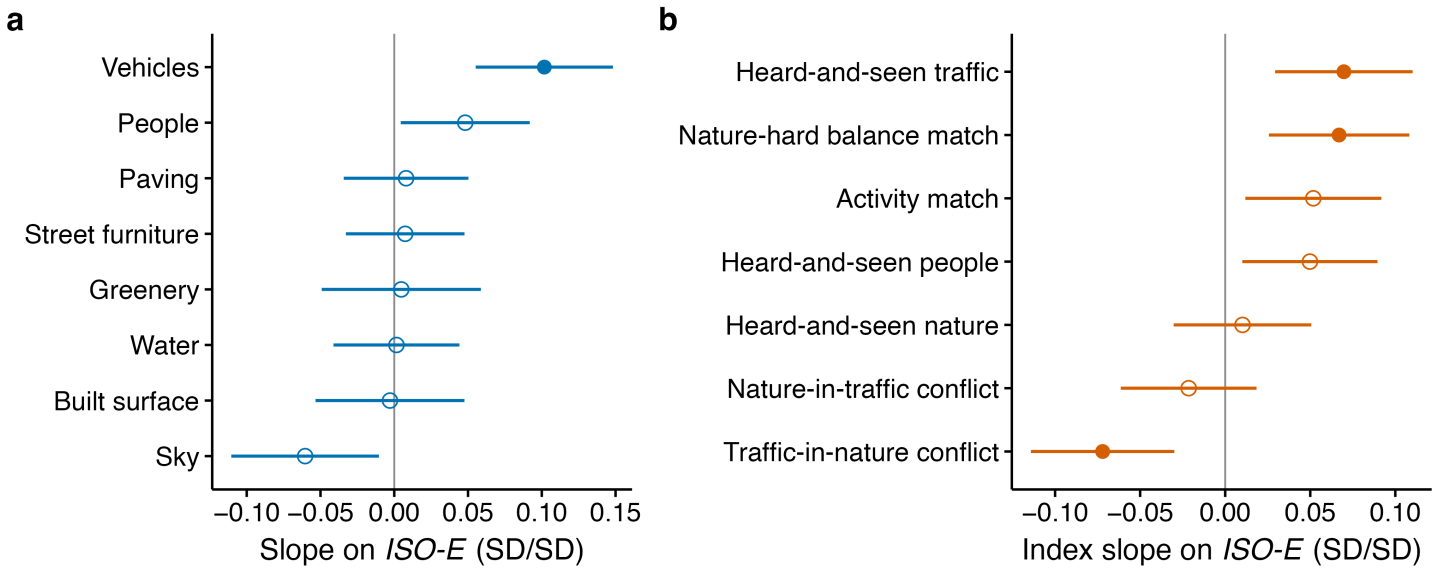


Figure 2

Eventfulness is the one evaluation dimension that scene composition predicts, alone and combined with perceived sound sources. (a) Slopes of the eight scene composites on eventfulness (ISO-E) from wave-adjusted siterandom- intercept linear mixed models (n = 2,262; points are estimates, bars 95% CIs). The filled point remains significant after Benjamini–Hochberg correction in the 40-test composition family ($q < 0.001$). (b) Slopes of the seven audio-visual match/conflict indices on ISO-E, each controlling its constituent main effects (same sample and model frame; filled points remain significant within the 35-test family, $q < 0.05$).

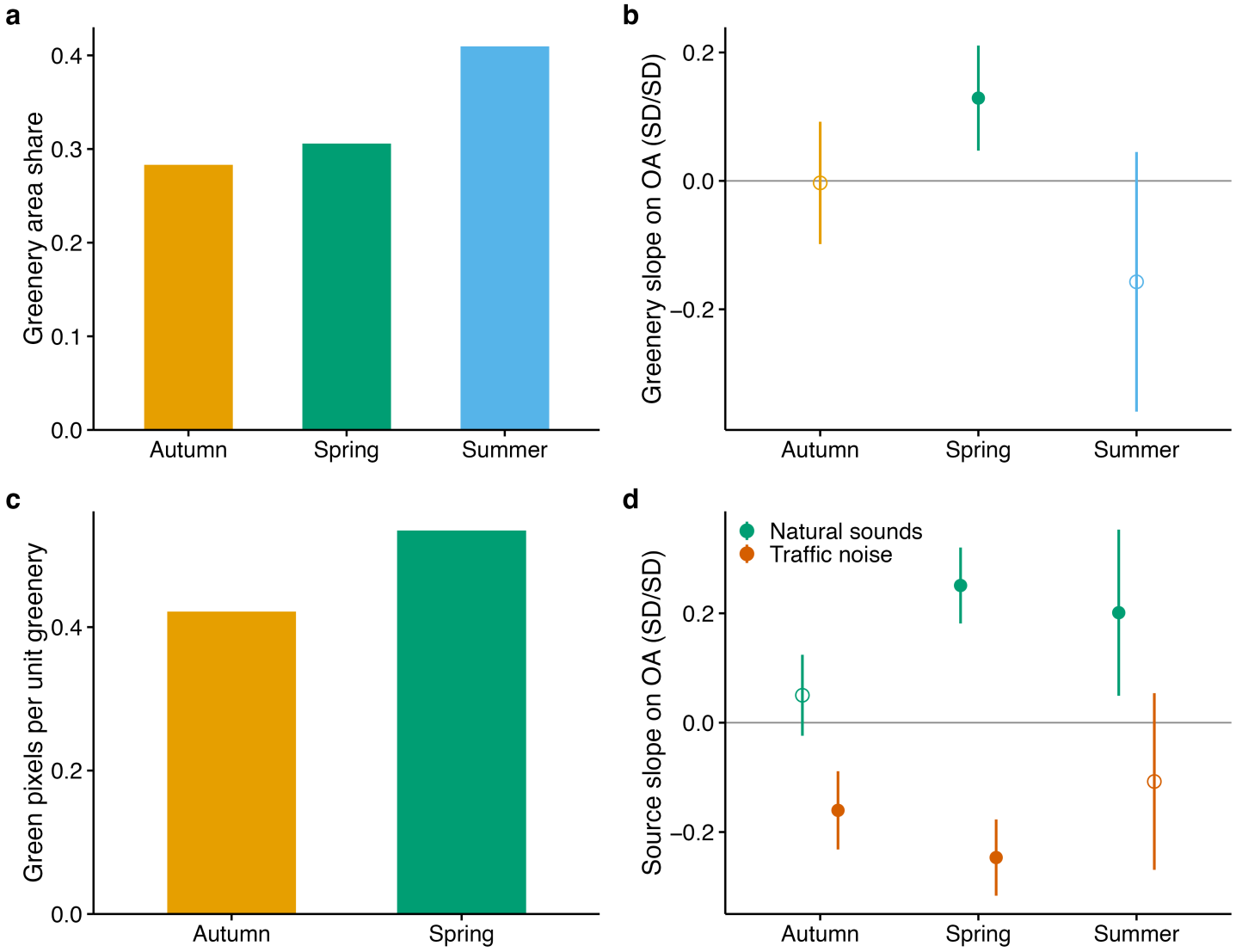


Figure 3

Greenery amount is constant between seasons while its condition and its association with evaluation change. All panels use the four multi-season sites (autumn 756, spring 851, summer 154 respondents), and the models in (b) and (d) have site and site-by-wave random effects. (a) Mean greenery areal share by season; the summer stratum is small and site-skewed. (b) Within-site greenery slope on overall evaluation (OA) by season (points are estimates, bars 95% CIs; filled points $p < 0.05$). (c) Green pixels per unit of segmented greenery by season, over the respondents whose greenery share exceeds 5% (n = 751 autumn, 756 spring photometric respondents). The autumn–spring contrast, from a model with a site random intercept and the capture-system proxy, holds under the site-cluster bootstrap (+0.11, 95% CI [0.07, 0.23]). (d) Within-site slopes of perceived natural sound and traffic noise on OA by season. The natural-by-season interaction remains significant within its 12- test family ($q = 0.018$), and the traffic slope shows no significant seasonal difference (joint $q = 0.157$).

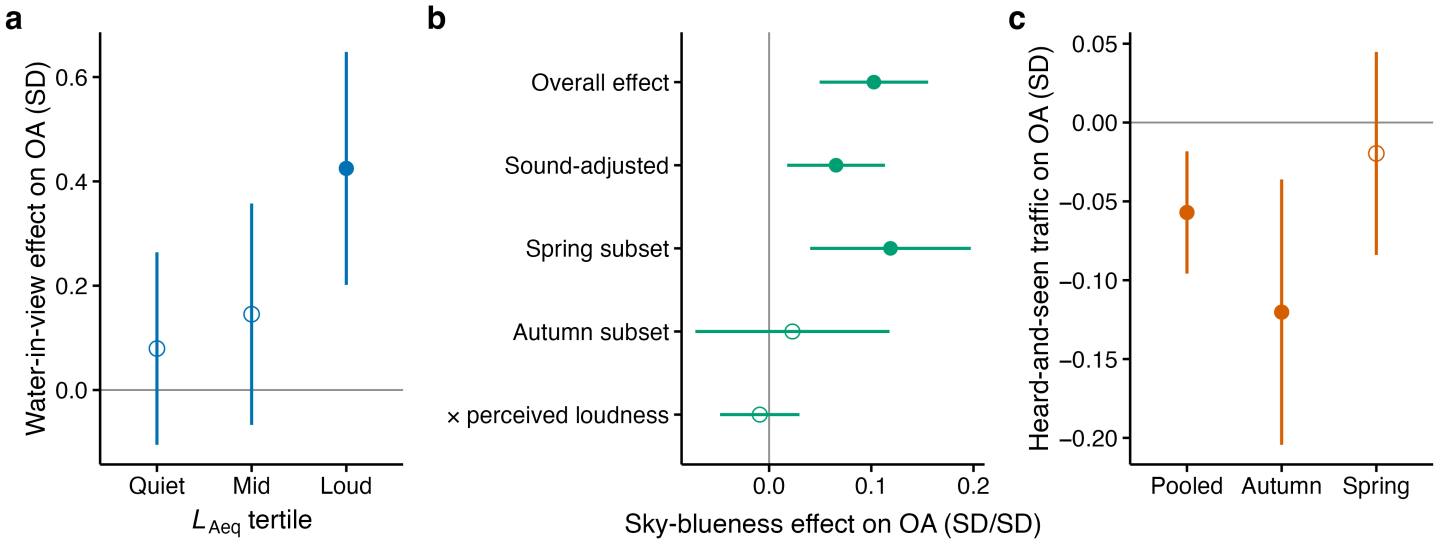


Figure 4

Visual associations across perceived sound characteristics. (a) Water-in-view effect on overall evaluation (OA) within LAeq tertiles of the photovalid acoustic sample (n = 743 per tertile; points are presence contrasts in outcome SD, bars 95% CIs, filled $p < 0.05$). The water-by-perceived-loudness interaction is +0.167 ($q = 0.010$ in the 7-test moderation family). (b) The sky-blueness effect on OA across specifications. The first is the condition-metric family estimate (n = 2,261; $q = 0.006$ in the 36-test family), and the second adds all auditory measures to the model (sound-adjusted). The others are the spring and autumn subsets and the interaction with perceived loudness (n = 2,228). All models adjust for sky share and the capturesystem proxy. The first two also include wave fixed effects and site and panorama-group random intercepts. The season subsets have site and site-by-wave random intercepts, and the loudness model adds wave fixed effects to these. (c) The heard-and-seen traffic interaction on OA, pooled (n = 2,262; $q = 0.036$ in the 9-test congruence family) and within the autumn and spring strata of the multi-season subset.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- supplementaryinformation.pdf