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.

Counting where it matters? Exposure coverage autocorrelation and hot-spot validity in network crash mapping

Diego Camargo

diegocamargo@cefetmg.br

Federal Center for Technological Education of Minas Gerais

Research Article

Keywords:

Posted Date: November 24th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11414338/v1

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

Additional Declarations: No competing interests reported.

Counting where it matters? Exposure coverage autocorrelation and hot-spot validity in network crash mapping

Diego Camargo[1*]

[1*]Department of Transportation Engineering, Federal University of Science and Innovation of Minas Gerais (UFCIMG), Belo Horizonte, MG, 30421-169, Brazil.

Corresponding author(s). E-mail(s): diegocamargo@cefetmg.br;

Abstract

Crash-rate maps guide road-safety investment, and their denominators come from traffic counting programs whose coverage is partial. This study investigates what happens to the map of crash-risk clusters when the coverage of the counting program is itself a spatial process. Using an analytical sample of 267,881 crashes assigned to panel bins over 8 years (2017–2020 and 2022–2025), drawn from 299,985 records read on the ten major Brazilian federal corridors with eligible monitored network, spatial autocorrelation of crash rates and coverage is measured at 5, 10 and 20 km, with false-discovery-rate (FDR) control on local statistics. Coverage is markedly more autocorrelated than risk (Moran's I of 0.780/0.661/0.431 for coverage versus 0.325/0.260/0.206 for rates), and their negative spatial association survives controls for traffic volume and sample precision with attenuation. In a misaligned pipeline that practitioners could run, two metropolitan high–high clusters survive FDR and make narrative sense, but are artefacts: implied volumes are impossible (1,164 vehicles/day in a metropolitan region of millions). With numerator and denominator aligned and coverage held constant, no high–high cluster survives FDR at the 10-km reference scale (70 nominal high–high clusters before correction; none after), while global structure persists. Without controlling for denominator geography, a crash-rate map risks inheriting the spatial footprint of the counting program: exposure coverage must be reported with any rate map, and multiplicity control is the entry condition for hot-spot claims.

Keywords: Road safety, Spatial autocorrelation, Hot-spot analysis, Exposure measurement error, False discovery rate, AADT

1 Introduction

Hot-spot maps are the operational instrument of road-safety management: they turn crash records into candidate sites for engineering intervention, and they do so at national scale. The statistical machinery behind them is mature, relying on kernel density and spatial autocorrelation statistics as well as clustering algorithms. Its uptake is documented across many national and municipal contexts [1–3]. One ingredient is not mature: the denominator. Crash *rate* maps normalize counts by exposure, and on highways exposure means traffic volume: the annual average daily traffic (AADT) produced by a counting program whose coverage of the network is partial and not arbitrary.

The exposure literature treats this partiality as a missing-data problem, with a clear split. One branch imputes: factor-based extrapolation, regression on infrastructure covariates, and machine-learning estimation of non-coverage AADT from open data [4]. Another branch propagates: recent work treats missing volume as a latent variable in a joint hierarchical model, so that exposure uncertainty flows into rate estimates [5]. Both improve the denominator; neither asks what the *shape* of non-coverage does to the map that the denominator serves.

This paper asks a structurally different question. If the geography of where volume is counted is itself spatially autocorrelated, then a rate map that does not control for it will absorb that geography. Environmental epidemiology formalized the impact of spatially structured measurement error when exposures predicted by spatial models are assigned to outcome locations [6–8]. This framework is transferred to road safety, distinguishing the exact mechanism: whereas unmonitored epidemiological exposures are typically predicted via kriging or land-use regression (inducing Berkson-like measurement error), the restriction of traffic data operates primarily as a spatially structured sampling bias and temporal misalignment of the denominator. The presence or absence of traffic monitoring is highly autocorrelated and dictates which segments have calculable rates, introducing a potential spatial confounder into the mapped risk.

The empirical evaluation is conducted on a real national program: 267,881 analytical crashes assigned to panel bins over 8 years (2017–2020 and 2022–2025), drawn from 299,985 records read on the ten major Brazilian federal corridors with eligible monitored network. The unit of analysis uses 1-km panel bins with stable federal jurisdiction and paved surface, assembled so that numerator and denominator always cover the same years. Global and local Moran statistics are computed on rates and on coverage across the 10 km reference scale as well as 5 and 20 km sensitivity cuts, with 9,999 permutations; local clusters must survive FDR control [9, 10] to be claims at all.

The paper's three contributions are:

- **A reframing.** Coverage of the exposure source is treated as a spatially structured confounder of the crash-rate map, not as missing volume. The transfer from environmental epidemiology is deliberate and explicit [6–8].
- **A measurement on a real program.** The spatial coupling between coverage and risk is quantified on a national counting program. This is founded by the direct cross-spatial association of the Bivariate Moran's I, which survives, with attenuation,

linear controls for traffic volume and sample precision, and by a counterfactual design. Coverage autocorrelation exceeds rate autocorrelation at every scale tested, and a misaligned pipeline manufactures FDR-surviving metropolitan clusters that implied-volume arithmetic exposes as artefacts.
- **A discipline.** Cluster claims are gated by FDR control on local statistics [9, 10] at the 10-km reference scale: 70 nominal high–high clusters before correction; none after. Bins failing this threshold do not reach statistical significance after false discovery rate control, remaining compatible with the null hypothesis of no local spatial autocorrelation given the multiplicity of tests.

The remainder of the paper is organized as follows. Section 2 reviews hot-spot analysis, exposure estimation, spatially structured measurement error, and multiplicity control. Section 3 details data and panel construction alongside the misalignment experiment. Section 4 reports results. Section 5 discusses implications and limitations, and Section 6 concludes.

2 Background

2.1 Hot-spot analysis in road safety

Identifying high-risk locations is the entry point of most road-safety management programs. GIS-based screening methods, notably spatial autocorrelation statistics and kernel density, are used to transform crash records into candidate sites for intervention, and their uptake has been documented across many national and municipal contexts [1–3]. Network screening typically ranks segments by crash frequency or crash *rate* [3]; the rate-based variant dominates practice because raw counts confound exposure with risk, and a busy, safe corridor can accumulate more crashes than a quiet, dangerous one. Local indicators of spatial association (LISA) [11] decompose global autocorrelation into per-location contributions and classify locations into spatial clusters (high–high and low–low) alongside mixed spatial outliers, producing the cluster maps that agencies read as hot-spot geography. LISA statistics have been applied to crash data at many scales, from city networks to whole countries [12–15]. However, systematic reviews of GIS and crash analysis indicate that while rate analysis is frequently used, hot-spot studies often rely on absolute frequencies or Kernel density without a flow denominator, or utilize static traffic volumes without interrogating the spatial coverage or temporal misalignment of the counting network [1–3]. These maps guide resource allocation, which is precisely why their validity conditions deserve scrutiny.

2.2 Exposure and its measurement

The denominator of a crash rate is exposure, and on highways exposure means volume. The reference metric is AADT, estimated from a mix of permanent and short-duration counting stations whose installation and maintenance are costly, so counting programs cover a fraction of the network directly and estimate the rest [4]. The estimation literature is rich: factor-based extrapolation and regression on infrastructure covariates, supplemented increasingly by machine-learning imputation of non-coverage AADT from openly available data [4]. In safety analysis, AADT enters crash prediction models

as the canonical exposure variable, and refinements disaggregate it temporally using continuous count stations and probe data [16]. Recent work goes further and treats missing volume as a latent variable in a joint hierarchical model, propagating exposure uncertainty into rate estimates [5]. What this literature shares is a framing of non-coverage as *missing information about volume*: a problem of estimation accuracy and uncertainty, to be solved with better imputers.

2.3 Spatially structured exposure error outside road safety

A parallel literature in environmental epidemiology asks a structurally different question: not *how large* the exposure error is, but *what shape* it has. When exposures are predicted from a rule or model and assigned to locations where outcomes are measured, the induced measurement error is not independent noise; it inherits the spatial correlation of the prediction process. Gryparis et al. characterized the impact of measurement error induced by spatially smoothed exposure surfaces [6]. Szpiro et al. formalized the decomposition of this error into classical and Berkson-like components arising from the spatial prediction of exposure at unmonitored locations [7]. Sheppard et al. consolidated the field's lesson, demonstrating that spatially correlated exposure measurement errors degrade statistical precision and introduce bias into spatial inference [8]. The mechanism in road safety, however, departs from the classical environmental epidemiology model. Whereas in epidemiology the exposure at unmonitored locations is predicted or smoothed via kriging or land-use regression (inducing Berkson-type error), in highway networks the data restriction operates primarily as a spatially structured sampling bias and a temporal denominator misalignment. The presence or absence of traffic monitoring is highly autocorrelated and dictates which segments have calculable rates and how the denominator is composed. This bias is substantiated directly through the Bivariate Moran's I, linking coverage to rates, and through a counterfactual design that contrasts cluster survival under varying versus constant exposure coverage.

2.4 Multiplicity in local spatial statistics

Local indicators invite a statistical hazard of their own: applying a per-location significance test across a network runs thousands of simultaneous tests, and nominal significance at each location inflates the family of false clusters. Caldas de Castro and Singer introduced FDR control for local statistics of spatial association, showing that the dependence structure of spatial tests requires care and that FDR-based corrections preserve far more power than family-wise alternatives [9]. The combined application of LISA statistics and FDR control has recently emerged as an established methodological precedent for spatial risk analysis at the municipal level [17]. For this study the operative lesson is twofold: cluster claims must survive FDR control to be claims at all, and the comparison of cluster maps before and after a correction must hold the multiplicity machinery fixed, so that differences trace to the correction's target and not to the testing regime.

2.5 Synthesis and gap

Three strands meet here. Hot-spot mapping supplies the applied instrument whose validity is at stake [1, 17]; the exposure literature supplies AADT and its estimation methods, but frames non-coverage as missing volume [4, 5, 16]; and spatially structured measurement-error theory supplies the mechanism: exposure assignment that shares spatial structure with the outcome biases the mapped result [6–8]. What is missing is the join: treating the coverage of the counting program as a spatial confounder of the crash-rate map, measured on a real program, with cluster inference disciplined by FDR control [10]. This paper makes that join.

3 Data and methods

The analytical framework developed in this study evaluates how the spatial structure of traffic counting programs affects network-wide crash risk mapping and local cluster inference. To avoid the conflation of risk with exposure availability, a longitudinal linear panel is constructed where crash occurrences and counting-station coverage are explicitly aligned across corridors and kilometre bands for each operational year. The overall study architecture is structured into four data stages (D1: raw sources covering PRF crashes and PNCT traffic counts alongside the SNV road inventory; D2: join and quality control; D3: corridor selection; D4: panel construction) and four analytical modules (M1–M4), illustrated schematically in Fig. 1.

3.1 Data (D1–D4)

Crashes come from the open records of the Polícia Rodoviária Federal (PRF), which polices, and therefore records on, the federal highway network exclusively. The study window is 2017–2025, complete years only. The corridor set is defined by exposure eligibility, not by geography alone: the ten federal corridors with the largest simulable vehicle-kilometres per day in the national counting program (BR-101, BR-116, BR-163, BR-153, BR-364, BR-381, BR-050, BR-376, BR-158, BR-135; approximately 28,300 km). The exact analytical attrition funnel begins with 299,985 crash records read during the 8-year study period. Finally, 32,104 records are excluded for falling outside the eligible network (lacking stable federal jurisdiction or unpaved sections). This exact sequence yields a final analytical sample of 267,881 crashes assigned to the panel bins. The overall study frame and panel dimensions, including data provenance, are summarized in Table 1.

Exposure comes from the annual average daily traffic (AADT) tables of the national counting program (PNCT/DNIT), joined to the road inventory (SNV) mesh. This program estimates the AADT through macroscopic network models that allocate origin-destination matrices from locations with known traffic volumes, calibrated by continuous and short-duration counting stations [18]. The join key is the physical reference BR plus kilometre mark, not segment identifiers, which the source re-issues across inventory versions; BR codes and decimal marks are normalized between sources. One corridor deserves explicit declaration regarding this eligibility criterion: BR-040 (Brasília–Belo Horizonte–Rio) is a major corridor but 53 of its 100 reference links are

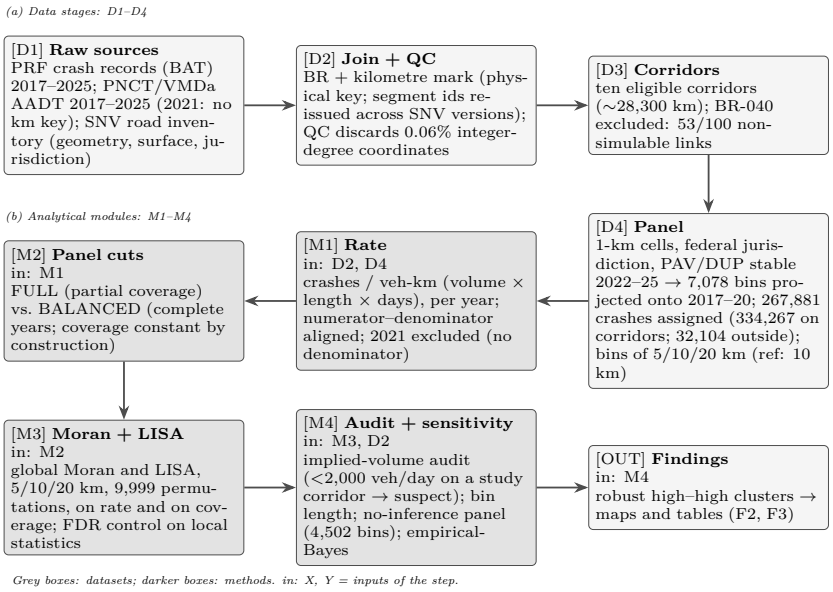

Fig. 1 Pipeline of the study. (a) Data: from raw sources D1–D4 to the analysis panel. (b) Methods: M1–M4, each stating its inputs (in: D/M). Grey boxes: datasets; darker boxes: methods. The rate (M1) is aligned with the denominator by construction; the audit (M4) tests whether surviving clusters are consistent with the exposure data.

classified non-simulable in the exposure source. This classification includes crossings without official traffic volume estimates, which prevents the calculation of contiguous exposure rates; its exclusion results directly from this documented cadastral eligibility criterion.

The counting network itself grew 66 per cent in modelled extent between 2017 and 2024, as the program expanded. This infrastructural expansion restricts valid temporal comparisons: a Moran coefficient computed per year is legitimate; comparing Moran coefficients across years on a changing mesh is not. Where the argument requires temporal statements, comparisons are restricted to the balanced panel defined in Section 3.6.

3.2 Units and the analysis panel

Crashes on highways are a linear phenomenon; areal units such as municipalities impose arbitrary boundaries while creating zero-denominator cells that break spatial contiguity. The unit of analysis is therefore the bin along each BR: contiguous bands of 10 km as the reference scale, with 5 and 20 km as declared sensitivity scales. Bin length is a linear variant of the modifiable-areal-unit problem [19]: there is no correct length, only the declared one and the test of whether patterns survive it; the same tension between zonal aggregation and estimated structure is documented for areal crash models [20–22].

Table 1 Study frame and panel sizes. Crash counts are national extractions of the ten selected corridors; panel sizes are bins on the projected panel (n varies with bin length).

Item	Value
Study window	2017–2025 (complete years)
Corridors (top ten eligible by simulable vehicle-km/day)	10 (BR-050, 101, 116, 135, 153, 158, 163, 364, 376, 381)
Corridor network length	∼28,300 km
Eligible network length (stable federal and paved)	22,304 km
Volume coverage across corridors	72–91 per cent
Records read against panel	299,985
Assigned to panel bins (analytical sample)	267,881
Full panel length (5 / 10 / 20 km)	22,135 / 22,203 / 22,279 km
Full panel bins (5 / 10 / 20 km)	4,540 / 2,329 / 1,206
Balanced panel length (5 / 10 / 20 km)	17,071 / 17,814 / 19,026 km
Balanced panel bins (5 / 10 / 20 km)	3,479 / 1,851 / 1,021
No-inference sensitivity panel	4,502 bins (31–34 per cent crash coverage)

Entry into the panel is by exposure, inverting the usual crash-first design. Formally, the ten selected corridors comprise approximately 28,300 km of registered length, of which 22,304 km meet the eligibility criteria of stable federal jurisdiction and paved or dual-carriageway surface (PAV/DUP) across all years that declare jurisdiction (2022–2025). These 22,304 km constitute the exact sum of the elemental 1-km panel bins (indexed by highway and state alongside kilometre marks) that meet these criteria, which are then projected retrospectively onto 2017–2020. These elemental cells are aggregated by state and corridor into analytical bands at the 10 km reference scale and 5 and 20 km sensitivity cuts. Because this aggregation generates residual bands at the extremities shorter than the nominal scale, multiplying the number of bins (4,540 bins at 5 km; 2,329 bins at 10 km; 1,206 bins at 20 km) by their nominal scale artificially exceeds the network length. The actual effective length with exposure in the full panel is the exact sum of the retained elemental bins: 22,135 km at 5 km (4,540 bins; 36 bands without exposure totalling 169 km are excluded), 22,203 km at 10 km (2,329 bins; 11 bands/101 km excluded), and 22,279 km at 20 km (1,206 bins; 2 bands/25 km excluded). The balanced panel, retaining only bins with usable volume data across all 8 complete years, represents 74.4 per cent of the eligible network's kilometres. In this balanced configuration, the effective retained length is 17,071 km at 5 km (3,479 bins), 17,814 km at 10 km (1,851 bins), and 19,026 km at 20 km (1,021 bins), which accounts for 76.5 per cent to 85.3 per cent of the eligible network. Three measured facts justify the projection. First, paved federal highways virtually never change custody (0.44 per cent turnover in four years). Second, while the aggregate crash capture of the ten corridors is 89.3 per cent (267,881 of 299,985 recorded crashes fall inside the eligible panel network, with 32,104 outside), the crash coverage inside the resulting panel is flat across the nine years (77.5–79.9 per cent, with no temporal drift when receding from 2025 to 2017). This stability validates the retrospective jurisdictional inference. Third, the alternative no-inference panel, explicit federal jurisdiction in every year, exists and is measured (4,502 bins, 31–34 per cent crash coverage) as a sensitivity cut. Jurisdiction matters on the numerator side as well: when a segment leaves the federal

network, the PRF stops recording on it, and a measured episode showed an 82 per cent crash drop on lost segments against a stable control, an artefact the stability rule removes by construction.

The year 2021 has no exposure key in the source (its AADT edition lacks kilometre bands). It is excluded from rate computation; no denominator is interpolated. Direction-resolved volumes are summed into a bidirectional exposure, a declared design choice that models each corridor as a single continuous topological axis and avoids directional disaggregation across undivided carriageways. Finally, to filter out noise from segments with near-zero traffic, exposure floor cuts are defined at the 10th and 25th percentiles. The panel without custody inference (4,502 bins) is also reported as a supporting sensitivity cut.

3.3 Rate definition (M1, M4)

The primary outcome of interest is the exposure-normalized crash rate R_i, defined per 100 million vehicle-kilometres for highway bin i following established highway safety engineering methodology [23]:

$$R_i = \frac{\sum_{t \in T_i} Y_{it}}{E_i} \times 10^8 \tag{1}$$

where Y_{it} denotes the observed crash count in bin i during operational year t, and T_i is the set of valid panel years. The cumulative denominator exposure E_i (expressed in vehicle-kilometres) is given by:

$$E_i = \sum_{t \in T_i} \mathrm{AADT}_{it} \cdot L_i \cdot D_t \tag{2}$$

where AADT_{it} is the bidirectional annual average daily traffic volume assigned to bin i in year t, L_i is the actual effective length of analytical band i (the sum of its eligible elemental bins with volume data, which prevents overestimating exposure in residual bands shorter than the nominal scales), and D_t is the number of calendar days in year t (365 or 366).

Numerator and denominator are temporally coupled by construction: T_i is identical for crash aggregation and exposure integration. When exposure is misaligned across years, the implied daily traffic volume diverges from physical highway reality. To systematically audit candidate clusters and verify denominator plausibility, the implied mean annual average daily traffic $\widehat{\mathrm{AADT}}_i$ is computed:

$$\widehat{\mathrm{AADT}}_i = \frac{E_i}{L_i \sum_{t \in T_i} D_t} \tag{3}$$

Under the proposed operational validation protocol, any bin on the top-ten federal corridors where $\widehat{\mathrm{AADT}}_i < 2{,}000$ vehicles/day is flagged as an exposure artefact, preventing spuriously inflated rates from contaminating downstream spatial models.

3.4 Spatial autocorrelation of coverage and of risk (M2)

To evaluate whether network-wide screening reflects true road safety hazards or the footprint of traffic monitoring, global spatial autocorrelation is measured symmetrically for both the crash rate R_i and the counting program's coverage C_i (defined as the number of usable exposure years per bin, $|T_i|$). Global spatial dependence is quantified using Moran's I statistic [11]:

$$I = \frac{n}{S_0} \frac{\sum_{i=1}^{n} \sum_{j=1}^{n} w_{ij}(z_i - \bar{z})(z_j - \bar{z})}{\sum_{i=1}^{n}(z_i - \bar{z})^2} \tag{4}$$

where $z_i \in \{R_i, C_i\}$ is the analyzed attribute for bin i, $\bar{z}$ is its sample mean across all n bins in the network panel, and w_{ij} are elements of the spatial weight matrix W. Because highway corridors operate as linear topological networks, W is defined by spatial contiguity along each corridor: $w_{ij} = 1$ if bins i and j share a contiguous boundary on the same highway axis, and $w_{ij} = 0$ otherwise. The scalar $S_0 = \sum_{i=1}^{n} \sum_{j=1}^{n} w_{ij}$ represents the sum of all network weights. Statistical significance is evaluated under 9,999 conditional random permutations. If $I(C) > I(R)$, the spatial distribution of data collection is more tightly clustered than crash risk itself, analogously reflecting how spatial structure in exposure assignment introduces spatially structured measurement error and distorts the mapped outcome, as conceptualized in the spatial exposure literature [6–8]. To ensure robustness against sample-size heterogeneity where sparse counts induce variance instability, rates are also evaluated under the Empirical Bayes spatial rate standardization of Assunção and Reis [24].

To ensure that observed rate autocorrelation reflects genuine spatial concentration of safety hazards rather than an artefact of the exposure denominator, two diagnostic benchmarks are specified. First, global Moran's I is computed univariately and symmetrically for the numerator (crash counts Y_i) and the denominator (exposure e_i). Second, a conditional null model is generated by simulating crashes as a Poisson process under spatially uniform risk proportional to traffic volume ($\lambda_i = e_i \times r$). Although crash counts typically exhibit overdispersion handled in the literature via negative binomial or Poisson-gamma formulations [25], the Poisson simulation acts strictly as a homogeneous spatial null benchmark. In 9,999 replications, empirical p-values are derived to test whether the observed rate dependence exceeds what would be produced by exposure structure alone.

Third, to formalize the built-in smoothness of the macroscopic network model, the fraction of neighbouring bin pairs sharing identical exposure from a common sublink is explicitly computed. This quantifies the structural redundancy across adjacent bins and provides the methodological justification for the mandatory control by false discovery rate and implied-volume audits.

3.5 Multiplicity control and local cluster inference (M3)

Local spatial clusters and spatial outliers are detected using Anselin's Local Indicator of Spatial Association (LISA) [11]. For each bin i, the local Moran statistic I_i is

computed as:

$$I_i = \frac{z_i - \bar{z}}{s_z^2} \sum_{j=1}^{n} w_{ij}(z_j - \bar{z}) \tag{5}$$

where $s_z^2 = \frac{1}{n}\sum_{i=1}^{n}(z_i - \bar{z})^2$ is the sample variance. Positive values of I_i denote spatial clustering of similar values (high–high hot spots or low–low cold spots), whereas negative values identify spatial outliers (high–low or low–high).

Because computing I_i entails simultaneous hypothesis testing across thousands of network segments ($m = n$), standard nominal significance thresholds ($p_i < 0.05$) generate substantial false discoveries. To discipline local cluster claims, all local pseudo-p-values obtained from 9,999 permutations are subjected to False Discovery Rate (FDR) control [9, 10]. Let $p_{(1)} \leq p_{(2)} \leq \cdots \leq p_{(m)}$ denote the ordered local p-values. The Benjamini–Hochberg FDR cut-off index k^* is determined by:

$$k^* = \max\left\{k \in \{1, \ldots, m\} : p_{(k)} \leq \frac{k}{m} q\right\} \tag{6}$$

where $q = 0.05$ represents the maximum tolerable proportion of false positives among rejected null hypotheses. Only bins satisfying $p_i \leq p_{(k^*)}$ are retained as statistically verified spatial clusters.

3.6 Panel cuts and reproducibility

The design contrast is structured around three distinct panel cuts, summarized in Table 2. The full panel admits bins with partial exposure coverage across the monitoring window, replicating standard engineering screening practice. The balanced panel (*anos_completos*) restricts inclusion to bins with valid counting records across all analyzed operational years, rendering coverage constant by construction and eliminating the spatial footprint of the counting program without parametric modeling. Finally, the no-inference panel provides a strict sensitivity baseline for readers unwilling to accept the retrospective projection of stable federal jurisdiction. Comparing global autocorrelation coefficients, local cluster survival before and after FDR, and continuous risk surfaces across these cuts directly isolates the confounding effect of exposure availability.

3.7 Computational implementation and reproducibility

All data processing, spatial indexing, exploratory spatial data analysis, and visualization were executed in a fully reproducible environment built on Python 3.12. Georeferencing, coordinate transformations (SIRGAS 2000, EPSG:4674), and geometric network segmentation were handled via GDAL/OGR and GEOS through GeoPandas 1.1.4 and Shapely 2.1.2. Linear network topology, spatial contiguity weight matrices (W), global Moran's I, empirical-Bayes rate standardization and local LISA statistics, paired with False Discovery Rate control (`fdr`) were computed using the PySAL ecosystem (`libpysal` 4.15.0 and `esda` 2.10.0). The local LISA statistic employs the empirical-Bayes standardized rate according to the formulation of Assunção and Reis [24], implemented in `esda.Moran_Local_Rate`. Panel construction and Monte

Table 2 Comparison of panel designs, inclusion rules, and analytical purposes across the ten highway corridors.

Panel cut	Inclusion rule	Bins (5/10/20 km)	Analytical purpose
Full panel	Usable exposure in ≥ 1 year	4,540 / 2,329 / 1,206	Baseline; mirrors unadjusted engineering practice
Balanced panel	Usable exposure across all analysed years	3,479 / 1,851 / 1,021	Eliminates monitoring footprint; holds exposure constant
No-inference	Annual explicit federal jurisdiction	4,502 (1-km bins)	Sensitivity baseline for retrospective jurisdictional projection

Carlo permutation tests were orchestrated with NumPy and SciPy. The 9,999 conditional Monte Carlo permutations were executed with a fixed pseudo-random seed (`seed=12345`) for both the global tests and the local `Moran_Local_Rate` indicators.

Reproducibility follows an end-to-end scripted discipline: bin construction, network adjacency graphs, panel inclusion cuts and permutation settings, ensuring that every empirical statistic reported in this manuscript trace directly to version-controlled pipeline artifacts in a public companion open-source repository of processed data and code (version-controlled via Git). This guarantees end-to-end reproducibility and aligns with the final declarations of data and code availability, providing open access and an auditable pipeline via the canonical repository (https://github.com/d-camargo/next-generation-road-safety). Publication figures and tables are generated deterministically from these audited outputs, eliminating undocumented manual post-processing steps.

4 Results

The empirical findings are organized around three complementary dimensions of spatial hot-spot validity across the Brazilian federal highway system. First, Section 4.1 quantifies the global spatial autocorrelation of counting program coverage alongside crash risk across multiple scales and panel designs, establishing that exposure availability possesses a substantially stronger spatial structure than the safety outcome it serves to normalize. Second, Section 4.2 analyzes the operational consequence of exposure misalignment, detailing how temporal mismatch in denominator construction manufactures spurious metropolitan clusters that survive multiplicity correction. Finally, Section 4.3 examines the local indicators of spatial association (LISA) and false-discovery-rate (FDR) control, demonstrating that once exposure is properly aligned and coverage variation is removed, nominal hot spots dissipate entirely while smooth spatial continuity remains.

4.1 Coverage is more autocorrelated than risk

To determine whether crash-rate screening reflects true risk patterns or the spatial imprint of data collection, global spatial autocorrelation of the counting program's

coverage is evaluated alongside that of the crash rate. Table 3 summarizes the resulting global Moran's I coefficients across the three aggregation scales (5, 10, and 20 km) for both the full and balanced panel designs. The comparative ordering between coverage and risk structure is illustrated in Fig. 2.

Table 3 Global Moran's I for coverage (usable exposure years per bin; full panel) and for the crash rate per vehicle-kilometre, by bin length. Rate coefficients are shown for both panel cuts. All coefficients are significant at $p \leq 0.001$ under 9,999 permutations, except the full-panel crash rate at 20 km ($p = 0.0084$ for the empirical-Bayes rate; $p = 0.0093$ for the raw rate).

Scale	n full	n balanced	I coverage	I rate (full)	I rate (balanced)
5 km	4,540	3,479	0.780	0.442	0.325
10 km	2,329	1,851	0.661	0.437	0.260
20 km	1,206	1,021	0.431	0.110	0.206

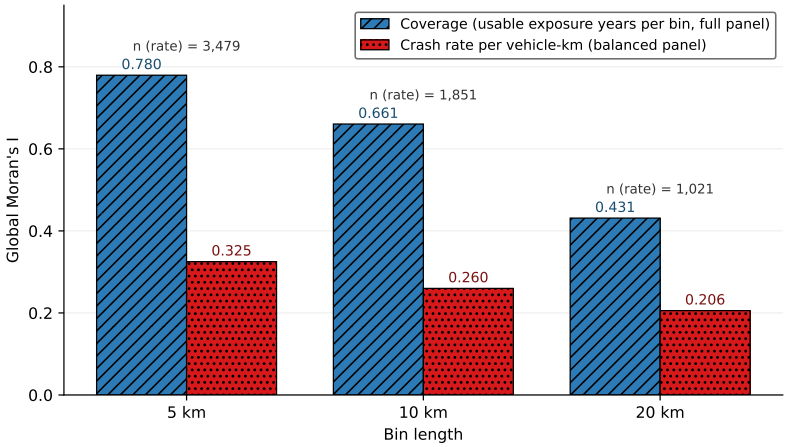

Fig. 2 Global Moran's I by bin length: coverage of the counting program (usable exposure years per bin, full panel) versus crash rate per vehicle-kilometre (balanced panel). Coverage structure exceeds rate structure at every scale, by a factor of 2.1–2.5. All coefficients are significant at $p \leq 0.001$ under 9,999 permutations.

As documented in Table 3 and Fig. 2, the central diagnostic is a measured *ordering* between the two spatial processes: the Moran's I values for coverage exceed those of the crash rate by a factor of 2.1–2.5 across all bin scales. This factor explicitly reflects the ratio between the coverage autocorrelation on the full panel (0.780, 0.661, and 0.431) and the crash rate on the balanced panel (0.325, 0.260, and 0.206). Measured as usable exposure years per bin on the full panel, coverage autocorrelation reaches $I = 0.780$ at 5 km, 0.661 at 10 km, and 0.431 at 20 km ($p = 0.0001$ across all scales). In contrast, the crash rate on the balanced panel exhibits substantially lower

autocorrelation: $I = 0.325$ at 5 km ($n = 3{,}479$), 0.260 at 10 km ($n = 1{,}851$), and 0.206 at 20 km ($n = 1{,}021$). Internal to the full panel, coverage autocorrelation exceeds that of the rate by factors of 1.8 at 5 km (0.780/0.442), 1.5 at 10 km (0.661/0.437), and 3.9 at 20 km (0.431/0.110). The full-panel rate I at 20 km (0.110) is smaller than the balanced-panel counterpart (0.206) because the inclusion of heterogeneous, low-count segments at the corridor ends attenuates contiguous spatial dependence, while the balanced panel isolates continuous stretches with stable annual coverage.

To examine the spatial structure of the denominator itself, a univariate decomposition of Moran's I at 10 km on the balanced panel of complete years contrasts the numerator of crashes ($I = 0.5077$, $p = 0.0001$) with the denominator of vehicle-kilometres ($I = 0.8407$, $p = 0.0001$). Testing the Poisson null hypothesis conditioned on exposure with 9,999 replications at the same scale and panel cut reveals that the mean simulated Moran's I under constant risk proportional to traffic is -0.0010 ($p = 0.0001$), compared to 0.2599 observed for the raw rate. This demonstrates that the null model reproduces less than 1 per cent (-0.40 per cent) of the observed spatial autocorrelation.

This empirical ordering rules out the uncritical reading of crash-rate maps as pure representations of safety risk. The footprint of the counting network is strongly clustered ($I = 0.661$ at the 10-km scale) because traffic counting stations are allocated following the institutional program design, which concentrates sensors along high-priority operational segments. On the full panel, the Bivariate Moran's I between coverage and risk is negative and significant ($I_{BV} = -0.276$, $p = 0.0001$, $z = -15.85$ at 10 km, and $I_{BV} = -0.330$, $p = 0.0001$, $z = -24.19$ at 5 km), supported by a statistically significant correlation ($r = -0.392$, $p < 10^{-85}$ at 10 km, and $r = -0.362$ at 5 km). When controlling for log(AADT), this negative association survives with slight attenuation: at 10 km ($n = 2{,}329$), the partial Pearson correlation shifts to -0.389 ($p < 10^{-84}$) and the partial Bivariate Moran's I to -0.267 ($p = 0.0001$, $z = -15.53$); at 5 km, Pearson shifts to -0.359 and Bivariate Moran's I to -0.322 ($p = 0.0001$). However, survival under linear control does not authorize treating coverage as a homogeneous spatial sampling bias over the entire network. Splitting the data by volume terciles reveals a lack of monotonicity: the Spearman correlation in the lowest traffic stratum (AADT $< 2{,}995$ at 10 km) is not statistically significant at 10 km ($r_s = 0.029$, $p = 0.418$) or 5 km ($r_s = 0.025$, $p = 0.332$). The negative association only emerges in higher volume strata ($r_s = -0.282$, $p < 10^{-14}$ at 10 km; $r_s = -0.309$ at 5 km in the third tercile), indicating that the phenomenon is strongly driven by the operational regime of high-traffic segments, which is consistent with sublinear exposure regimes documented in the literature [5]. Precision control via Empirical Bayes standardization (Assunção–Reis) practically does not shrink the rates due to the large accumulated volume (median weights $w \geq 0.98$). However, directly testing the hypothesis of sample dispersion reveals that while the rate variance indeed decreases monotonically with coverage (at 10 km, the variance drops from 253,721 in bins with 1–3 years of coverage to 3,429 in bins with 7–9 years), the negative association survives weighting by exposure (proportional to the inverse of the theoretical Poisson variance) with attenuation: the exposure-weighted Pearson correlation is -0.279 at 5 km, -0.275 at 10 km, and -0.246 at 20 km (all $p < 0.001$). Therefore, sample precision does not

fully explain the association between coverage and risk. Within coverage strata, however, the association is weak and inconsistent across measures: in the 7–9-year class at 10 km ($n = 2{,}071$), where coverage takes only the values 7 and 8, Pearson is -0.070 ($p = 0.001$) and Spearman -0.035 ($p = 0.112$). Dispersion that shrinks with coverage is compatible with sampling imprecision but does not establish it, and surviving the weighting does not identify the cause of the association. Despite these concurrent mechanisms, this contrast provides empirical justification for the paper's two central design choices: adopting the balanced panel, which holds coverage constant by construction, and implementing an implied-volume audit for candidate clusters.

4.2 Spurious clusters that survive FDR until the exposure is aligned

In the full panel (the counterfactual at 10 km, prior to balancing), three high–high clusters survive strict FDR control. These comprise three contiguous bins on BR-364/AC (km 110–140) with intermittent coverage. The drop to zero high–high clusters on the balanced panel results from the removal of denominator intermittency. Auditing by implied volume (implied AADT = crashes / (rate $\times$ km $\times$ days $\times 10^{-8}$)) demonstrates that on these intermittent segments, the implied volume fell artificially below 2,000 vehicles/day, revealing the underestimation of the denominator. Conversely, in monitored conurbated segments, the implied volume exceeds 20,000 vehicles/day. The spatial impact of this misalignment extends beyond local peaks to the global structure: the temporally misaligned execution artificially shifted the global Moran's I at 10 km to 0.364. When evaluated under strict temporal alignment, the autocorrelation is 0.437 on the full panel and 0.260 on the balanced panel, demonstrating how denominator mismatch distorts both global dependence and local cluster detection.

The diagnostic audit of implied volumes detected this failure. To enforce denominator plausibility across the panel, a strict traffic filtering rule is applied: any candidate bin whose implied temporal coupling falls below the floor of 2,000 vehicles/day is systematically flagged as an exposure artefact and discarded, given that actual flows in metropolitan approaches typically exceed tens of thousands of vehicles per day. The exposure assigned to the emblematic case of BR-101 in Pernambuco (km 70–80), the primary arterial gateway of the Recife metropolitan area, failed this plausibility floor. In the temporally misaligned pipeline, eight years of crash records (1,448 occurrences) were calculated against an exposure denominator derived from only the single usable counting year (2017). This arithmetic mismatch implied a fictitious average daily traffic volume of just 1,164 vehicles/day, forging a spurious high–high cluster that survived FDR. Once numerator and denominator years were strictly aligned, restricting the analysis to the single year with the denominator, the crash count for that bin dropped to 156. The implied volume adjusted to a credible 9,310 vehicles/day, matching the observed counting records, and the nominal cluster dissolved below the FDR threshold. This case demonstrates that multiplicity correction alone cannot protect screening against structural exposure misalignment.

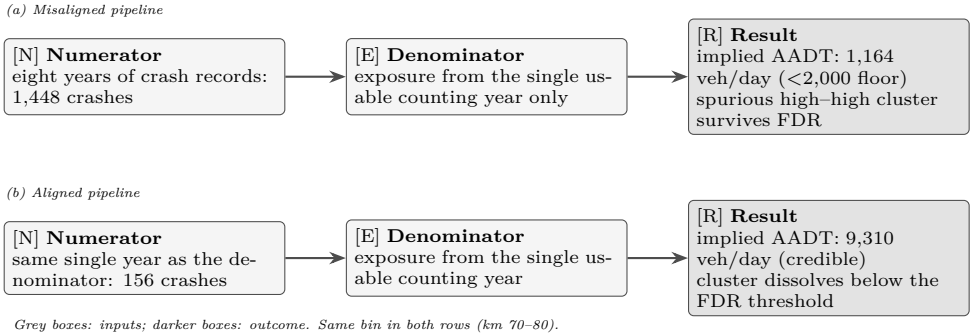

Fig. 3 Misaligned versus aligned pipeline for the emblematic bin (km 70–80, Recife metropolitan approach). N: crash count (numerator); E: exposure (denominator); R: result. AADT: annual average daily traffic, here implied by crashes and exposure; FDR: false discovery rate. (a) Eight years of crashes (1,448) against the exposure of a single counting year imply 1,164 veh/day, below the 2,000 veh/day floor, and a spurious high–high cluster survives FDR. (b) With numerator and denominator in the same year (156 crashes), the implied volume is 9,310 veh/day and the cluster dissolves. Grey boxes: inputs; darker boxes: outcome.

4.3 What survives: global structure, no continuous clusters

Having removed coverage variation through the balanced-panel design and eliminated temporal misalignment, local spatial association is examined. The identification of critical highway safety investments relies on whether hot spots emerge as statistically distinct local concentrations or whether spatial dependence reflects smooth regional gradients. Figure 4 displays the LISA cluster map for the balanced network, comparing nominal associations against FDR survivors, while Fig. 5 illustrates the corresponding continuous crash-rate distribution across the ten corridors.

Examination of Figs. 4 and 5 reveals three structural properties of network-level highway risk. First, at the uncorrected nominal threshold ($p < 0.05$), the network produces 70 apparent high–high clusters, clustered predominantly near large urban centres. These nominal clusters span twenty-five highway sections (nine federal highways) across eighteen states, heavily concentrated in Minas Gerais (12 bins), Paraná (11 bins), and Rio de Janeiro (9 bins). Concurrently, nominal low–low clusters dominate extensive inter-urban stretches, totalling 202 bins.

Second, under false-discovery-rate control ($q \leq 0.05$), all high–high clusters at the 10 km reference scale are rejected (70 nominal high–high clusters before correction; none after). At the other scales, exactly 1 high–high cluster survives FDR (1 of 159 at 5 km; 1 of 22 at 20 km on the balanced panel). This sole surviving high–high cluster is located on BR-101/SC (km 130–135 and km 120–140, the urban/conurbated crossing of Balneário Camboriú and Itajaí). It is characterized by very high actual flow and an observed count exceeding 2,000 crashes, with an implied volume consistent with the capacity of a dual-carriageway highway. This survival confirms the sensitivity of the spatial scale and justifies the adoption of the 10 km reference scale, ensuring conclusions are not bound to assumptions refuted by a linear MAUP effect. As indicated

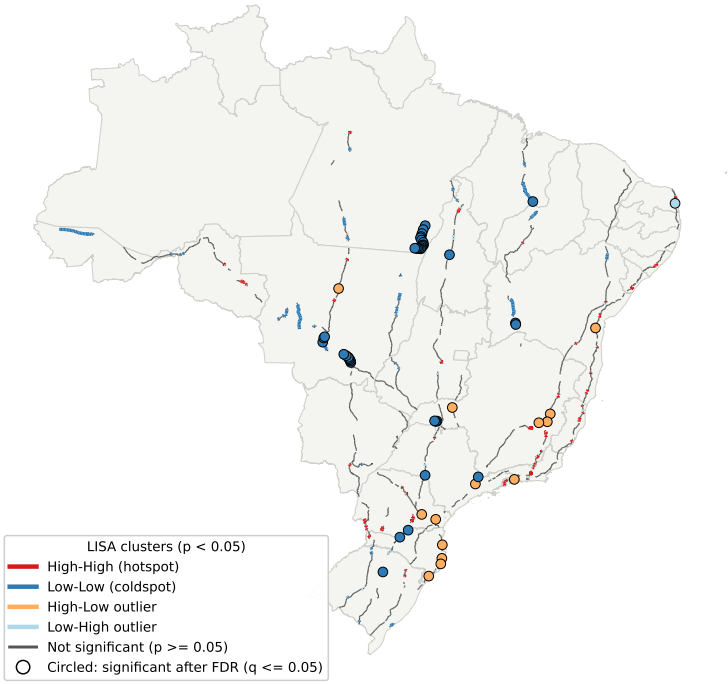

Fig. 4 LISA cluster map of the crash rate (10 km bins, balanced panel, $n = 1{,}851$), at the nominal threshold $p < 0.05$. Significant-after-FDR bins ($q \leq 0.05$) are marked with black-edged markers. Quadrants: high–high (hotspots), low–low (coldspots), high–low and low–high outliers. No high–high bin survives FDR at the 10 km scale; surviving bins are low–low and spatial outliers.

by the highlighted markers in Fig. 4, the absence of contiguous high–high clusters does not imply the absence of local peaks. The only localized high-risk features that withstand multiplicity control at the reference scale are 14 isolated high–low spatial outliers, surviving as punctual local peaks significantly higher than their surroundings. A single low–high outlier also survives on BR-101/RN (km 110–120), situated directly adjacent to an extreme local rate peak (764 crashes per 100 million vehicle-km at km 100–110).

Third, the continuous rate surface in Fig. 5 maps the variation of crash risk across the network. The 75th percentile across the balanced panel is 50 crashes per 100 million vehicle-km. Isolated extreme values occur sparsely across all ten corridors, notably on BR-101/RN (km 100), BR-364/RO (km 230 and 420), BR-116/RS (km 40), BR-153/TO (km 140), and BR-163/MT (km 830), representing local point anomalies rather than coherent regional clusters.

Table 4 details the local cluster counts across the three aggregation scales before and after FDR discipline.

On the reference cut (10 km, balanced panel, 9,999 permutations), the global rate autocorrelation is positive and significant ($I = 0.260$, $p = 0.0001$) and robust to empirical-Bayes rate adjustment. The twelve combinations represent the parametric

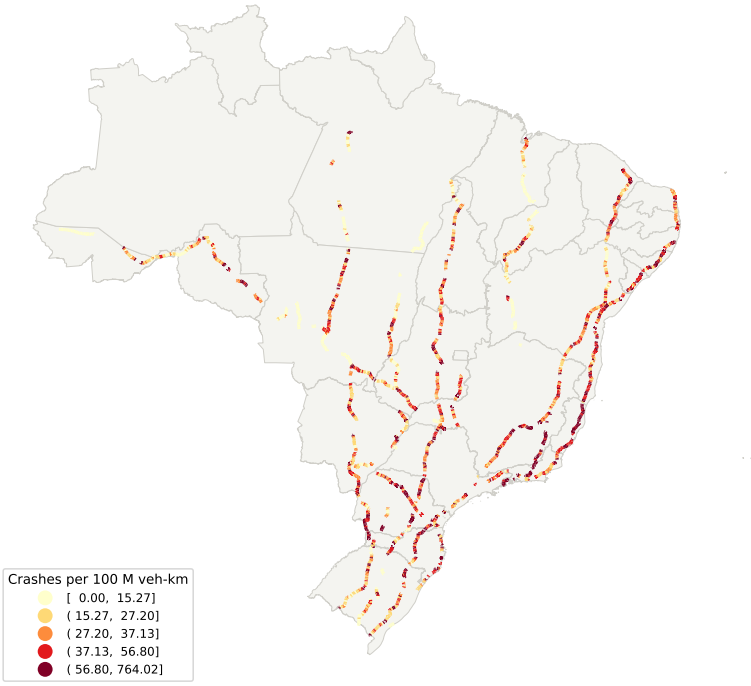

Fig. 5 Crash rate per 100 million vehicle-kilometres along the ten corridors (10 km bins, balanced panel), in five quantile classes.

matrix resulting from the 3 spatial scales (5 km, 10 km, 20 km) crossed with 4 exposure sensitivity cuts (full panel without floor, floor at the 10th percentile, floor at the 25th percentile, and balanced panel with 8 full years). In all 12 combinations, the raw and empirical-Bayes standardized rates agree in sign and statistical significance, as detailed in Table 5 (all with $p \leq 0.0093$ under 9,999 permutations with `seed=12345`, with coefficients varying from $I = 0.106$ to 0.550). At the local level, the count of surviving rejections closes exactly with k^* under the Benjamini–Hochberg procedure. At the 10 km reference scale (balanced panel, $n = 1{,}851$, FDR cut-off 1.4×10^{-3}), exactly $k^* = 53$ bins survive, comprising 0 high–high clusters, 38 low–low clusters, along with 15 spatial outliers (14 high–low and 1 low–high). These 14 high–low outliers, together with one low–high outlier located on BR-101/RN (km 100–110), total 15 spatial outliers that sum with the 38 low–low bins to complete the 53 survivors.

This arithmetic closure holds across all scales: at 5 km (balanced panel, $n = 3{,}479$, cut-off 6.5×10^{-4}), exactly $k^* = 45$ bins survive (1 HH, 33 LL, 11 HL, 0 LH); at 20 km (balanced panel, $n = 1{,}021$, cut-off 1.8×10^{-3}), exactly $k^* = 37$ bins survive (1 HH, 28 LL, 8 HL, 0 LH). The counterfactual explains the collapse of nominal hot spots: under variable coverage (full panel), the spatial sampling bias of the counting program (quantified by the Bivariate Moran's I ($I_{BV} = -0.276$, $p = 0.0001$, $z = -15.85$ at 10 km, and $I_{BV} = -0.330$, $p = 0.0001$, $z = -24.19$ at 5 km) and Pearson correlation

Table 4 Local cluster counts before and after FDR control, by scale and panel cut. The final column reports the critical cut-off threshold $p_{(k^*)} \leq (k^*/m)q$ dictated by the exact Benjamini–Hochberg closure rule (Eq. 6). At the reference scale (10 km) every nominal high–high cluster is rejected on the balanced panel.

Scale	Panel	n	HH pre	HH post	LL pre	LL post	HL post	LH post	k^*	FDR cut-off
5 km	Full	4,540	207	3	570	36	4	0	43	4.7×10^{-4}
5 km	Balanced	3,479	159	1	419	33	11	0	45	6.5×10^{-4}
10 km	Full	2,329	96	3	267	17	1	0	21	4.5×10^{-4}
10 km	Balanced	1,851	70	0	202	38	14	1	53	1.4×10^{-3}
20 km	Full	1,206	31	2	143	18	0	0	20	8.3×10^{-4}
20 km	Balanced	1,021	22	1	112	28	8	0	37	1.8×10^{-3}

Table 5 Sensitivity matrix of the global Moran's I for the crash rate across combinations of spatial scale and panel exposure cut. The table reports both the raw and empirical-Bayes (EB) standardized coefficients, with their respective simulated p-values.

Scale	Panel cut	n	I (raw)	p (raw)	I (EB)	p (EB)
5 km	Full	4,540	0.406	0.0001	0.442	0.0001
5 km	p10	4,087	0.550	0.0001	0.550	0.0001
5 km	p25	3,405	0.507	0.0001	0.507	0.0001
5 km	Balanced	3,479	0.325	0.0001	0.325	0.0001
10 km	Full	2,329	0.434	0.0001	0.437	0.0001
10 km	p10	2,096	0.299	0.0001	0.300	0.0001
10 km	p25	1,747	0.295	0.0001	0.295	0.0001
10 km	Balanced	1,851	0.260	0.0001	0.260	0.0001
20 km	Full	1,206	0.106	0.0093	0.110	0.0084
20 km	p10	1,085	0.186	0.0005	0.187	0.0005
20 km	p25	904	0.239	0.0001	0.240	0.0001
20 km	Balanced	1,021	0.205	0.0003	0.206	0.0004

($r = -0.392$, $p < 10^{-85}$ at 10 km, and $r = -0.362$ at 5 km) allows 3 high–high clusters to survive FDR. By fixing coverage as constant on the balanced panel, the bias is removed, and zero high–high clusters survive.

These results characterize the empirical nature of spatial risk: there is global spatial structure in risk, but no continuous clusters distinguishable from their neighbours once multiplicity is disciplined. Risk varies smoothly along the corridors, with extreme hazard concentrating strictly in punctual, isolated peaks rather than cartographically detachable contiguous zones. A cluster that fails FDR is not necessarily a weak hot spot; it does not reach statistical significance after false discovery rate control ($q \leq 0.05$), being compatible with the null hypothesis of no local spatial autocorrelation given the multiplicity of tests.

The reference scale of 10 km is a design choice with three declared grounds. It is coarse enough to accumulate exposure in sparsely counted stretches: at 10 km the balanced panel retains 1,851 bins against 3,479 at 5 km, yet keeps the crash counts per bin high enough for rate stability (the empirical-Bayes adjustment that smooths raw rates matters most where counts are thin). It is fine enough to separate metropolitan approaches from inter-city stretches: at 20 km, corridor ends merge with their metropolitan rings and the local statistics lose resolution (global I drops from 0.260 to 0.206 while local clusters coarsen). And it is the scale at which the two panels (full and balanced) remain directly comparable, since both keep the same corridor network at 10 km with coverage stable across years. The 5 and 20 km scales are reported as declared sensitivity cuts, not as competing designs: the ordering of coverage over risk autocorrelation holds at all three (Fig. 2).

Three sensitivity analyses corroborate these findings. The first varies bin length across the 10 km reference and the 5 and 20 km sensitivity scales; global structure remains positive and significant across all three (Table 3). The assertion regarding the spatial dependence of coverage holds consistently: monitoring coverage is substantially more autocorrelated than the crash rate across all panel cuts where temporal coverage varies (full panel without floor, as well as p10 and p25 exposure floors). In the balanced panel, however, coverage is constant by design (exactly 8 years for each bin), rendering coverage autocorrelation structurally inapplicable or null. The second uses the no-inference panel, composed of 4,502 elementary 1 km bins with strictly declared jurisdiction in all years. This panel verifies the institutional stability of jurisdiction (31–34 per cent crash coverage), suggesting that the retrospective projection did not forge contiguity artefacts. The third analysis is the metropolitan-approach audit of implied volumes for the surviving high–high clusters. Across all evaluations, the qualitative ordering survives, and no high–high cluster survives FDR control at the reference scale once exposure is aligned.

5 Discussion

The central finding is an ordering: the geography of the counting program is more autocorrelated than the geography of the risk it is meant to normalize. Coverage Moran's I of 0.780/0.661/0.431 against rate I of 0.325/0.260/0.206 at 5/10/20 km means that any rate map produced without coverage control partially maps the footprint of the counting network. In the misaligned program design this was not a theoretical risk: it manufactured two metropolitan high–high clusters that survived FDR control and made narrative sense, until implied-volume arithmetic exposed them. While echoing the spatially structured measurement error formalized in environmental epidemiology [6–8], the mechanism here manifests as a persistent empirical association, and stratification by volume indicates that direct mechanical causality remains open to competing explanations of sublinear exposure regimes (safety-in-numbers) [5].

5.1 Implications for practice

Rate-based network screening inherits coverage structure wherever exposure is partial, and the practical consequences are immediate. First, exposure coverage must be

reported alongside any crash-rate map, the way response rates are reported alongside survey estimates: a map of rates without a map of where the denominator comes from is incomplete evidence, and the reporting costs nothing; the coverage surface is already computed as a by-product of the join between crashes and counting stations. Second, the balanced-panel discipline, restricting claims to bins whose exposure history is complete, is a cheap and defensible alternative to modeling the coverage process when the analytic question tolerates the restriction; it costs spatial coverage of the analysis (1,851 of 2,329 bins at 10 km), and that cost must be reported too. Third, the implied-volume audit generalizes: any surviving hot-spot cluster can be checked for plausibility through the denominator it implies, and incorporating that arithmetic into routine hot-spot QA is recommended: it requires no additional data, only the reverse computation from rate and count back to vehicles per day, compared against what the corridor plausibly carries. Together the three measures turn coverage from an invisible property of the data into a reported and systematically audited design dimension.

5.2 Relation to the literature

The imputation literature addresses missing volume as missing information [4], and the uncertainty-propagation literature treats unobserved exposure as a latent variable whose uncertainty flows into rate estimates [5]. Both improve the denominator; neither treats the coverage of the counting program as a spatial confounder of the mapped risk. The hot-spot literature, meanwhile, applies LISA and cluster detection to crash data with growing sophistication. Systematic reviews demonstrate that traffic-weighted rate analysis is a canonical method in road safety screening [1–3], yet the literature frequently assumes the denominator as a stable given, without interrogating the spatial autocorrelation of the monitoring presence itself as a potential confounder. The contribution of this study slots precisely into that join: rather than refining the imputation of missing volumes, it establishes operational constraints for the resulting map, with multiplicity control as the gatekeeper of cluster claims [9, 10].

5.3 Limitations

Five are declared. (1) State-boundary fragmentation: adjacency is built within (BR, state) chains because kilometre marks restart at each state line, so the BR-101 is eleven separate chains and no bin pair crosses a border; stitching requires inventory geometry, not kilometre marks, a local effect on clusters near borders, declared and left as future work. (2) Bidirectional exposure: volumes are summed over directions; direction-resolved analysis doubles the units and is future work. (3) Jurisdictional projection: the panel defines the network by stable federal jurisdiction over 2022–2025 and projects it back to 2017–2020; three measured facts support the projection (0.44 per cent custody turnover on paved federal highways; flat 77.5–79.9 per cent crash coverage across nine years; the 4,502-bin no-inference panel as a sensitivity cut), but it remains an inference, and readers unwilling to accept it have the alternative panel reported. (4) No year of exposure for 2021: rates are not interpolated; the year is excluded from rate computation rather than assigned an invented denominator. (5)

Built-in smoothness of the macroscopic network model: because the AADT estimation relies on network flow allocation from known count locations [18], at 10 km on the sample without an exposure floor, neighbouring pairs possess identical exposure and share a sub-link in at least one common year, justifying the mandatory control by FDR and implied-volume audit.

6 Conclusions

This study investigated what happens to the map of road-crash risk clusters when the coverage of exposure, the counting program that produces the AADT, is itself a spatial process more autocorrelated than the risk it serves to estimate. On Brazilian federal highways, the answer is that the map inherits the geography of the counting program. Coverage autocorrelation exceeds rate autocorrelation at every scale tested (0.780/0.661/0.431 against 0.325/0.260/0.206 at 5/10/20 km) and this negative association between coverage and risk persists with attenuation under linear volume and precision controls, though competing non-linear exposure mechanisms remain viable explanations; a misaligned pipeline manufactured metropolitan clusters that survived multiplicity correction; and once coverage structure is held constant by construction, the local hot-spot picture collapses at the 10-km reference scale: 70 nominal high–high clusters before correction; none after, while global spatial structure remains.

The practical translation: a crash-rate map is only as trustworthy as the geography of its denominator. Agencies should report exposure coverage alongside rate maps, audit surviving clusters through their implied volumes, and treat multiplicity control as the entry condition for any hot-spot claim. The theoretical translation: the transfer from environmental epidemiology is conceptual: framing coverage not as a mere missing-data nuisance, but as a spatially structured covariate with potential confounding effects on rate estimation. However, instead of importing corrections based on calibration regression or parameter bootstrap of kriging-predicted surfaces (inapplicable to the sampling bias and temporal misalignment mechanism of road networks) [7], traffic engineering must adopt its own operational correction discipline: balanced panel designs with stable coverage, systematic multiplicity control via FDR, supplemented by implied-volume plausibility audits.

Future work: stitching corridors across state boundaries with inventory geometry; direction-resolved exposure; exposure surfaces inferred with propagated uncertainty [5]; and extending the coverage-as-confounder framing to other rate-based maps in transportation analytics.

Declarations

- **Ethics approval and consent to participate:** Not applicable. This study is based exclusively on publicly available, anonymized administrative data from government agencies (Federal Highway Police and DNIT); in accordance with national regulations and institutional policies, formal ethics committee approval and participant consent are not required.
- **Funding:** This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors.

- **Conflict of interest:** The author declares no conflict of interest.
- **Data availability:** Crash data are publicly available from the Federal Highway Police (PRF) open-data portal. Traffic exposure data are publicly available from the National Department of Transport Infrastructure (DNIT). Derived panel datasets including network adjacency graphs and spatial weights are made openly available at https://github.com/d-camargo/next-generation-road-safety and permanently archived via Zenodo (DOI: 10.5281/zenodo.23167885).
- **Code availability:** The complete analytical and replication code for all models, permutation tests, tables, and figures is available at https://github.com/d-camargo/next-generation-road-safety.
- **AI usage:** During the preparation of this work, the author used Large Language Models (generative AI) for analytical script structuring, pipeline automation, and language polishing. The author critically reviewed all generated content and takes full responsibility for the integrity and conclusions of the publication.

References

[1] Endashaw, D.K., Habtegiorgis, K.A., Al-Ramadan, B.M., Al-Ahmadi, H.M., Deressa, B.F.: A systematic review on GIS-based road traffic accidents analysis and road safety audit. Computational Urban Science **5**(1) (2025) https://doi.org/10.1007/s43762-025-00221-w

[2] Skaug, L., Nojoumian, M., Dang, N., Yap, A.: Road Crash Analysis and Modeling: A Systematic Review of Methods, Data, and Emerging Technologies. Applied Sciences **15**(13), 7115 (2025) https://doi.org/10.3390/app15137115

[3] Ziakopoulos, A., Yannis, G.: A review of spatial approaches in road safety. Accident Analysis & Prevention **135**, 105323 (2020) https://doi.org/10.1016/j.aap.2019.105323

[4] Taiwo, A.A., Shuaib, B., Comert, G., Begashaw, N., Adjei Owusu, J.: Improved non-coverage annual average daily traffic estimation using openly available infrastructure data. Digital Engineering **10**, 100113 (2026) https://doi.org/10.1016/j.dte.2026.100113

[5] Skaug, L.: Beyond Empirical Bayes: A Hierarchical Bayesian Approach to Crash Rate Estimation with Missing Traffic Volume. Preprint, arXiv (May 2026). https://doi.org/10.48550/arXiv.2605.27889 . https://arxiv.org/abs/2605.27889

[6] Gryparis, A., Paciorek, C.J., Zeka, A., Schwartz, J., Coull, B.A.: Measurement error caused by spatial misalignment in environmental epidemiology. Biostatistics **10**(2), 258–274 (2008) https://doi.org/10.1093/biostatistics/kxn033

[7] Szpiro, A.A., Sheppard, L., Lumley, T.: Efficient measurement error correction with spatially misaligned data. Biostatistics **12**(4), 610–623 (2011) https://doi.org/10.1093/biostatistics/kxq083

[8] Sheppard, L., Burnett, R.T., Szpiro, A.A., Kim, S.-Y., Jerrett, M., Pope, C.A., Brunekreef, B.: Confounding and exposure measurement error in air pollution epidemiology. Air Quality, Atmosphere & Health **5**(2), 203–216 (2011) https://doi.org/10.1007/s11869-011-0140-9

[9] Castro, M., Singer, B.H.: Controlling the False Discovery Rate: A New Application to Account for Multiple and Dependent Tests in Local Statistics of Spatial Association. Geographical Analysis **38**(2), 180–208 (2006) https://doi.org/10.1111/j.0016-7363.2006.00682.x

[10] He, Z., Liu, Q., Deng, M., Xu, F.: Handling multiple testing in local statistics of spatial association by controlling the False Discovery Rate: A comparative analysis, pp. 684–687. IEEE, ??? (2017). https://doi.org/10.1109/icbda.2017.8078722 . http://dx.doi.org/10.1109/icbda.2017.8078722

[11] Anselin, L.: Local Indicators of Spatial Association—LISA. Geographical Analysis **27**(2), 93–115 (1995) https://doi.org/10.1111/j.1538-4632.1995.tb00338.x

[12] Gedamu, W.T., Plank-Wiedenbeck, U., Wodajo, B.T.: A spatial autocorrelation analysis of road traffic crash by severity using Moran's I spatial statistics: A comparative study of Addis Ababa and Berlin cities. Accident Analysis & Prevention **200**, 107535 (2024) https://doi.org/10.1016/j.aap.2024.107535

[13] Alam, M.S., Tabassum, N.J.: Spatial pattern identification and crash severity analysis of road traffic crash hot spots in Ohio. Heliyon **9**(5), 16303 (2023) https://doi.org/10.1016/j.heliyon.2023.e16303

[14] Ouni, F., Belloumi, M.: Pattern of road traffic crash hot zones versus probable hot zones in Tunisia: A geospatial analysis. Accident Analysis & Prevention **128**, 185–196 (2019) https://doi.org/10.1016/j.aap.2019.04.008

[15] Hazaymeh, K., Almagbile, A., Alomari, A.H.: Spatiotemporal Analysis of Traffic Accidents Hotspots Based on Geospatial Techniques. ISPRS International Journal of Geo-Information **11**(4), 260 (2022) https://doi.org/10.3390/ijgi11040260

[16] Dutta, N., Fontaine, M.D.: Improving freeway segment crash prediction models by including disaggregate speed data from different sources. Accident Analysis & Prevention **132**, 105253 (2019) https://doi.org/10.1016/j.aap.2019.07.029

[17] González-Sánchez, E.J., Pasqualino, J., Ayala-García, J.: Spatial Dependence and Cluster Persistence in Drinking-Water Quality Risk: Municipal Evidence from the Department of Santander, Colombia, 2007–2020. Water **18**(17), 2089 (2026) https://doi.org/10.3390/w18172089

[18] Silva, R.S.d., Santos, I.V., Cunha, F.C.S.: Síntese da metodologia para estimativa do Volume Médio Diário Anual - VMDa 2025 em toda a malha rodoviária federal pavimentada. Nota Técnica 001/2026, DNIT / Consórcio Concremat/Certare,

Brasília (January 2026). Contrato PP-831/2020-00 - Lote 5

[19] Cai, Q., Abdel-Aty, M., Lee, J., Eluru, N.: Comparative analysis of zonal systems for macro-level crash modeling. Journal of Safety Research **61**, 157–166 (2017) https://doi.org/10.1016/j.jsr.2017.02.018

[20] Wang, S., Chen, Y., Huang, J., Chen, N., Lu, Y.: Macrolevel Traffic Crash Analysis: A Spatial Econometric Model Approach. Mathematical Problems in Engineering **2019**(1) (2019) https://doi.org/10.1155/2019/5306247

[21] Sipos, T., Afework Mekonnen, A., Szabó, Z.: Spatial Econometric Analysis of Road Traffic Crashes. Sustainability **13**(5), 2492 (2021) https://doi.org/10.3390/su13052492

[22] Dong, N., Huang, H., Zheng, L.: Support vector machine in crash prediction at the level of traffic analysis zones: Assessing the spatial proximity effects. Accident Analysis & Prevention **82**, 192–198 (2015) https://doi.org/10.1016/j.aap.2015.05.018

[23] American Association of State Highway and Transportation Officials: Highway Safety Manual, 1st edn. AASHTO, Washington, DC (2010)

[24] Assunção, R.M., Reis, E.A.: A new proposal to adjust Moran's I for population density. Statistics in Medicine **18**(16), 2147–2162 (1999) https://doi.org/10.1002/(sici)1097-0258(19990830)18:16⟨2147::aid-sim179⟩3.0.co;2-i

[25] Washington, S.P., Karlaftis, M.G., Mannering, F.: Statistical and Econometric Methods for Transportation Data Analysis. Chapman and Hall/CRC, ??? (2003). https://doi.org/10.1201/9780203497111 . http://dx.doi.org/10.1201/9780203497111