Systematic biases in mobile phone mobility data from heterogeneous tower density

Leo Ferres[1,2*] and Erick Elejalde[3*]

[1]Faculty of Engineering, Universidad del Desarrollo, Santiago, Chile.
[2]ISI Foundation, Turin, Italy.
[3]L3S Research Center, Leibniz University Hannover, Hannover, Germany.

*Corresponding author(s). E-mail(s): lferres@udd.cl; elejalde@l3s.uni-hannover.de;

Abstract

Mobile phone data have become a cornerstone of human mobility research, offering unprecedented spatial and temporal coverage at population scale. However, the spatial distribution of cell towers introduces systematic biases that are rarely quantified. Urban towers are typically spaced hundreds of metres apart, while rural towers may be separated by several kilometres. This heterogeneity creates a spatially structured measurement floor: short trips in low-density areas are invisible, rural populations are misattributed, and origin–destination matrices are artificially urban-centric. In this paper, we develop a six-step correction pipeline and apply it to the Región Metropolitana of Santiago, Chile, using a 63,832-antenna catalog, 2024 census data, and eXtended Detail Records (XDR) from a major telecom operator. We show that replacing the standard tower-point Voronoi tessellation with sector-aware polygons yields a 3.0× spatial resolution gain, that intra-site sector transitions recover 100 million additional short-distance trips (median displacement 429 m) invisible at the tower level, and that inverse probability weighting uplifts rural comuna flows by 50–73%. A Fay–Herriot small area model provides additional smoothing, with shrinkage weights ranging from $\gamma \approx \mathbf{0.7}$ in the urban core to $\gamma < \mathbf{0.1}$ at the periphery. The pipeline is implemented in the open-source Python library `mobilens`, which requires only a tower catalog, a census population layer, and a study area boundary.

Keywords: Mobile phone data, Tower density bias, Inverse probability weighting, Small area estimation, Human mobility, H3 hexagonal grid, Sector geometry

1 Introduction

Mobile phone records, including call detail records (CDRs) and the higher-resolution eXtended Detail Records (XDRs) generated by data traffic, have been extensively used to study human mobility patterns, from daily commuting and urban activity to epidemic spreading and disaster response [1, 5, 10, 12, 16, 17, 19, 20]. Their appeal lies in the combination of large population coverage, fine temporal granularity, and longitudinal availability, which together offer a portrait of collective movement that no travel survey can match [2, 22]. Yet the measurement apparatus that produces these data is not spatially uniform. Telecom operators deploy towers according to demand and revenue, concentrating infrastructure in cities and leaving rural areas with sparse coverage. This asymmetry has long been acknowledged in the literature, but its quantitative consequences for mobility estimation are seldom characterised in detail [9].

The core of the problem is that mobile phone positioning is *tower-relative*: a user is located not by their true coordinates but by the identity of the tower they are connected to. Standard practice assigns each user to the Voronoi cell of their serving tower, implicitly equating the tower position with the user's location [9, 15]. In the city, where towers in Santiago can be as close as 22 m apart

(after merging co-located sites on the same mast), the resulting Voronoi cells are small and the approximation is reasonable. In the rural periphery of the Región Metropolitana, where inter-tower distances reach 14.7 km, a single Voronoi cell may cover an area larger than a small municipality. The spatial precision of the observation is thus orders of magnitude coarser in rural areas than in urban ones.

This heterogeneity gives rise to three interlocking biases. First, *movement invisibility*: a trip shorter than the local inter-tower distance never crosses a cell boundary and is therefore unobservable, creating a spatially varying detection floor that truncates the short-distance tail of the displacement distribution. Second, *population misattribution*: when users are assigned to large rural Voronoi cells, population density computed as users per cell area implicitly assumes a uniform distribution within the polygon, an assumption that is wildly wrong when a single cell contains both a town centre and 50 km^2 of farmland. Third, *urban-centric origin–destination matrices*: because both detection probability and phone penetration rate correlate with tower density, raw OD flows systematically undercount rural mobility relative to urban mobility. Critically, all three biases are spatially structured and correlated with urbanicity, the very variable that researchers most often wish to study [9].

Several approaches to mitigating these biases have been proposed [4, 6, 8, 13, 15, 23]. Census-calibrated ratio estimation rescales mobile phone user counts to match census population totals, correcting the level but not the spatial distribution of the error. Dasymetric redistribution uses ancillary data such as land cover, building footprints, or road networks to redistribute population within Voronoi cells more realistically [4, 15]. Inverse probability weighting (IPW) treats each observed trip as a sample drawn with a spatially varying inclusion probability and reweights accordingly [6, 13]. Small area estimation techniques such as the Fay–Herriot model borrow strength from structural models (e.g. gravity) to stabilise noisy estimates in data-sparse regions [8, 23]. However, these methods have typically been applied in isolation, and few studies have exploited the directional information available in modern tower catalogs, in particular the antenna azimuth and height, to refine the spatial model of tower coverage beyond the isotropic Voronoi assumption [11].

In this paper, we develop an integrated correction pipeline that chains six steps: (1) sector geometry inference from antenna azimuth and height, (2) detection floor characterisation, (3) dasymetric census redistribution to an H3 hexagonal grid, (4) OD matrix construction with intra-site mobility recovery, (5) inverse probability weighting, and (6) Fay–Herriot small area estimation. We apply the pipeline to the Región Metropolitana de Santiago (R13), Chile, using a 63,832-antenna tower catalog, the 2024 national census, and XDR mobility traces from a 6.5-week observation window. The pipeline is implemented in the open-source Python library `mobilens`, designed to be operator- and country-agnostic: it requires only a tower catalog with azimuth and height, a census population layer at any administrative level, and a study area boundary polygon.

2 Data

2.1 Tower catalog

Our primary data source is the 2025 antenna catalog from a major Chilean telecom operator, containing 63,832 antenna records nationwide. After restricting to the Región Metropolitana and removing records with missing coordinates, we retain 22,836 cells corresponding to 2,600 base transceiver station (BTS) sites. Each record includes a cell identifier, a BTS identifier (grouping co-located cells), geographic coordinates, antenna height above ground, beam azimuth, and radio technology. The technology breakdown in R13 reflects a mature network: 1,192 cells on 2G, 9,502 on 3G, 9,996 on 4G, and 2,146 on 5G. Antenna heights range from 0 to 91 m with a median of 24 m, and the distribution is visibly bimodal, with one mode around 3 m (indoor small cells and repeaters) and another around 30 m (standard macro towers).

2.2 Census cartography

We use the 2024 Chilean census for R13, which provides population counts at the *manzana* (city block) level, the finest available spatial unit. The data comprise 66,873 manzanas with a combined population of 7,059,040 inhabitants [21]. Median population per manzana is 64 persons; 16,476 manzanas (25%) are uninhabited, corresponding to parks, industrial zones, and undeveloped land. The cartography is distributed as a GeoPackage in SIRGAS 2000 (EPSG:4674).

2.3 XDR mobility traces

eXtended Detail Records (XDR) from the same operator cover the period from 1 August to 14 September 2023, a 6.5-week window selected because it falls between the winter school break and the Fiestas Patrias holiday and is therefore free of major calendar disruptions. This "boring" mobility baseline is precisely what is needed for characterising the tower-density bias structure rather than seasonal or event-driven anomalies. The raw cell-to-cell transition table contains 25.9 million OD pairs representing 931 million trips. Users are filtered to those observed at a minimum of two distinct BTS within R13 during the period, excluding stationary devices (IoT sensors, modems, alarm systems).

2.4 Road network

To support the dasymetric population redistribution, we download the OpenStreetMap drivable road network for R13 using the `osmnx` library [3]. The resulting network comprises 375,478 road segments and is cached locally for reuse.

3 Methods

Our correction pipeline consists of six steps, each building on the outputs of the previous ones. The first three steps (sector geometry, detection floor, and census redistribution) require only the tower catalog and census data; the remaining three (OD construction, IPW, and Fay–Herriot) additionally require the XDR traces.

3.1 Sector geometry inference

The standard approach in the literature assigns each mobile phone user to the Voronoi cell of their serving tower, treating each tower as an isotropic point [9, 15]. In practice, most towers are sectorised: they carry multiple directional antennas, each serving a wedge-shaped coverage area defined by the antenna's azimuth and beamwidth. We exploit this directional information to construct a finer-grained spatial model.

Infrastructure classification.

The bimodal height distribution suggests two distinct classes of infrastructure. Antennas with height below 10 m are indoor small cells, street-level repeaters, or distributed antenna systems that add network capacity but do not define spatial coverage regions in the conventional sense. We classify these as *micro* cells (4,333 cells, 19% of the R13 catalog) and exclude them from the spatial analysis. The remaining *macro* cells (18,503, 81%) form the basis of the sector tessellation.

Co-location merging.

Many BTS identifiers in the catalog correspond to equipment co-located on the same physical mast, representing different technology generations or network functions that share a single structure. To avoid artificially inflating the number of sites and producing zero-distance nearest-neighbour pairs, we cluster BTS positions within 20 m of each other using complete-linkage hierarchical clustering. This reduces 2,600 BTS to 1,292 physical sites; of these, 813 BTS were found to share a mast with at least one other BTS and were merged into co-located groups, while the remaining sites had a single BTS each.

Sector grouping.

At each physical site, multiple technologies (2G, 3G, 4G, 5G) may share the same physical antenna and thus the same azimuth. We round azimuths to the nearest 5° and group macro cells by (site, rounded azimuth) to identify physical sectors, i.e. distinct directional antennas. This yields 5,378 sectors across 1,292 sites. The dominant configurations are tri-sector (428 sites, 33%), six-sector (255 sites, 20%), two-sector (133 sites, 10%), and omni-directional (66 sites, 5%).

Beamwidth and coverage radius.

For each sector, beamwidth is inferred as $360°/n_s$, where n_s is the number of sectors at the site. Coverage radius is first estimated from antenna height using the radio-horizon formula $r = \sqrt{2hR_e}$, where $R_e \approx 8{,}500$ km is the effective Earth radius accounting for atmospheric refraction. This formula

gives a theoretical maximum range that depends only on height, not on the proximity of neighbouring towers. In practice, however, urban cells are interference-limited rather than range-limited, so we cap the radius at $0.65 \times d_{\text{nn}}$, the nearest-neighbour inter-site distance, which makes the effective radius sensitive to local tower density. This cap is binding for approximately 95% of sites; only the most isolated rural towers approach their radio horizon. The median calibrated radius is 367 m.

Sector polygons and centroids.

Each sector is represented geometrically as a circular wedge centred on the tower position, with the calibrated radius and inferred beamwidth. All geometry is computed in UTM 19S and stored in SIRGAS 2000 for compatibility with the census cartography. Sector centroids are offset from the tower along the azimuth direction at distance $(2r/3)\sin(\alpha/2)/(\alpha/2)$, the centroid of a circular sector with half-angle $\alpha/2$ and radius r. The median centroid offset is 224 m.

The sector tessellation yields a median effective resolution of 299 m, compared with 894 m for the conventional tower-point Voronoi, a 3.0× gain in spatial precision (Figures 1 and 2).

3.2 Detection floor

A trip originating in the coverage area of tower k is only observed if the user crosses a sector boundary or leaves the tower's coverage area entirely. The probability of detection depends on the trip displacement relative to the local spatial resolution. We define the effective resolution of site k as $r_{\text{eff},k} = r_k/n_{s,k}$, derived from the arc width of a sector wedge: at the coverage boundary the arc is $2\pi r_k/n_{s,k}$, and $r_{\text{eff},k}$ retains the proportional part $r_k/n_{s,k}$ (the constant 2π is absorbed by the exponential scale parameter below). More sectors produce narrower wedges, reducing the lateral distance a user must travel to cross a sector boundary. We model the detection probability as a function of displacement Δx (great-circle distance between origin and destination):

$$\delta_k(\Delta x) = 1 - \exp\left(-\frac{\Delta x}{r_{\text{eff},k}}\right) \tag{1}$$

This functional form arises naturally from the assumption that the user starts at a uniformly random position within the sector and moves in a uniformly random direction. The 50% detection threshold, defined as the trip length above which detection becomes more likely than not, is then $d_{50} = r_{\text{eff}} \cdot \ln 2$.

Across the 1,292 sites in R13, the effective resolution ranges from 1.3 m to 3,667 m, with a median of 91 m and a corresponding median d_{50} of 63 m. At the 95th percentile of the distribution (representing rural and peri-urban sites), r_{eff} reaches 578 m and d_{50} reaches 401 m; the most isolated site in the region has a detection floor of 2,542 m (Table 1). These numbers quantify the spatial scale below which mobility is largely invisible to the network, though individual trips shorter than r_{eff} can still be detected if they happen to cross a sector or site boundary (Figure 3; see also Figure 4 for the full distribution of d_{50}).

We validate the parametric model with a Monte Carlo simulation. We select three example sites at the 5th, 50th, and 95th percentiles of the r_{eff} distribution to represent the urban, suburban, and rural extremes. For each site we draw 20,000 random trips per distance value and comparing the simulated detection rate with the exponential prediction. The simulated rates consistently exceed the parametric values, especially at suburban and rural sites, because the single-tower model does not account for the additional detection opportunity provided by neighbouring towers (Figure 8). This conservative bias is by design: the parametric formula provides a lower bound on detection probability, which is the safe direction for the inverse probability weighting that follows.

3.3 Census redistribution to H3 grid

To provide a spatially explicit population denominator for the penetration rate estimation, we redistribute census population from manzana boundaries to a uniform H3 hexagonal grid at resolution 8, where each cell covers approximately 0.74 km^2. We redistribute census population from manzana boundaries to an H3 resolution-8 grid. Polyfilling the R13 boundary yields 24,258 hexagons. The administrative region itself has an area of about 15,402 km^2, and 2,517 hexagons (10.4%) contain population, reflecting the concentration of settlement in the urban footprint while the Andean foothills and peripheral valleys remain largely uninhabited.

The baseline redistribution uses standard areal interpolation: each manzana's population is allocated to overlapping hexagons in proportion to the intersection area [4]. We refine this with two

additional spatial weights: (i) sector coverage, defined as the number of tower sectors from the tessellation that overlap each hex (a proxy for the presence of population, since operators site towers where people are), and (ii) road-network density, defined as the total length of OSM drivable roads within each hex. Both weights are multiplied with the area weight and renormalised within each manzana, so that the total population is exactly conserved ($\sum = 7{,}059{,}040$ across all three methods).

The sector-weighted redistribution shifts population toward tower-covered hexagons and away from uncovered ones (mean absolute shift 21 persons per hex, maximum 3,012), with the largest shifts concentrated at the urban–rural transition where manzanas are large enough to straddle hexagons with different tower coverage.

3.4 OD matrix with intra-site correction

We construct the origin–destination matrix by joining the cell-to-cell transition table from BigQuery with the tower catalog to attach site identity, antenna azimuth, and coverage radius to each transition. The join matches 93.7% of transitions, corresponding to 889 million of the original 931 million trips; the unmatched records arise from cells present in the XDR data but absent from the 2025 catalog (likely decommissioned cells, tower catalogues are cummulative over the years).

Intra-site transitions.

A notable feature of the data is that transitions between different cells at the same physical site (intra-site transitions) account for 354 million trips, or 39.8% of all matched transitions. These would be entirely invisible in a tower-level OD matrix. We distinguish two types of intra-site transitions:

- **Sector crossings** (100 million trips): the user was handed off to a sector with a different azimuth at the same site, indicating real spatial movement around the tower. We estimate the displacement from the angular separation between sectors: $\hat{d} = (2r/3) \cdot 2\sin(|\Delta\theta|/2)$, where r is the coverage radius and $\Delta\theta$ the azimuth difference. The median estimated displacement is 429 m.
- **Technology handoffs** (254 million trips): the user remained on the same azimuth but was switched between technologies (e.g. from 3G to 4G). These transitions carry no spatial information and are excluded from the displacement analysis.

Including the sector crossings enriches the displacement distribution in the 100–500 m range, which is entirely absent from inter-site-only data (Figure 5). The inter-site median displacement is 5,855 m, consistent with typical urban commuting distances in Santiago reported in the EOD 2012 origin–destination survey [18]. The combined OD matrix is aggregated to 658,539 H3-level pairs and 2,692 comuna-level pairs.

3.5 Inverse probability weighting

The observed trip counts are a biased sample of true mobility: trips in low-density areas are under-detected and populations in those areas may be under-represented among phone users. We model the overall inclusion probability (the chance that a trip both originates from a phone user in our data *and* is spatially detected) for a trip of length Δx originating at site k as:

$$p_k(\Delta x) = \hat{\rho}_k \cdot \delta_k(\Delta x) \tag{2}$$

where $\hat{\rho}_k = \min(U_k/N_k, 1)$ is the estimated phone penetration rate (observed unique users U_k divided by census population N_k in the site's H3 cell, capped at 1 to accommodate commuters and multi-SIM devices), and δ_k is the detection probability from Equation 1. Because we have data from only a single operator, the market-share component is absorbed into $\hat{\rho}_k$ and assumed spatially uniform.

The raw IPW weight $w = 1/p$ is stabilised using the Hájek estimator (normalised so that $\bar{w} = 1$) to prevent a small number of high-weight rural trips from dominating the corrected totals [6, 13]. In practice, the resulting weights are moderate: 75% of trips receive a weight of 0.86 (a slight urban downweight), and only 0.2% of trips exceed a weight of 2. The inclusion probability is floored at 0.01 to bound the maximum weight.

The IPW correction produces a substantial uplift in flows originating from peripheral comunas (Table 2, Figure 6). The five most affected are Alhué (1.73×), María Pinto (1.72×), San Pedro (1.72×), Pirque (1.51×), and Curacaví (1.45×), all located in the southwestern or southern periphery of R13, where tower density is lowest. Urban core comunas are correspondingly downweighted by the Hájek normalisation (∼0.86×).

Table 1 Detection floor across R13

Site type	r_{eff} (m)	d_{50} (m)	Interpretation
Urban (5th pctl)	24	16	Sub-block trips (∼50 m) detected ∼88%
Suburban (50th pctl)	91	63	Trips under 63 m are coin-flips
Rural (95th pctl)	578	401	Trips under 400 m usually missed
Most isolated	3,667	2,542	Trips under 2.5 km invisible

Table 2 IPW uplift for the five most affected comunas

Comuna	Raw trips	IPW trips	Uplift
Alhué (13505)	619,853	1,069,917	1.73×
María Pinto (13504)	1,141,676	1,962,161	1.72×
San Pedro (13303)	675,562	1,159,734	1.72×
Pirque (13103)	6,121,013	9,266,686	1.51×
Curacaví (13503)	4,369,481	6,317,713	1.45×

Table 3 Three-way comparison for selected comunas

Comuna	Raw trips	IPW trips	FH trips	γ
Santiago (urban)	73.9M	71.8M	71.4M	0.71
Maipú (suburban)	31.4M	27.9M	27.4M	0.58
Pirque (rural)	6.1M	9.3M	9.5M	0.28

3.6 Fay–Herriot small area estimation

While the IPW correction adjusts the expected value of the flow estimates, it does so at the cost of increased variance, particularly for rural comuna pairs where few trips are observed and the weights are large. To reduce this variance, we apply a Fay–Herriot model that borrows strength from a structural gravity specification [8, 23, 24].

The model treats the log-transformed IPW-corrected flow for each comuna pair (i, j) as a direct estimate with known sampling variance:

$$\hat{F}_{ij}^{\text{IPW}} = \mathbf{x}_{ij}'\boldsymbol{\beta} + u_{ij} + e_{ij} \tag{3}$$

where $\mathbf{x}_{ij}$ is a covariate vector comprising log population of origin and destination, log inter-centroid distance, and a self-loop indicator; $\text{Var}(e_{ij}) = \psi_{ij}$ is the sampling variance, modelled as the mean effective resolution of the pair; and $u_{ij} \sim N(0, \sigma_u^2)$ is the model-level random effect, with σ_u^2 estimated by method of moments.

All four covariates are highly significant ($p < 0.001$). The coefficients are consistent with a standard gravity model: flows increase with population at origin ($\hat{\beta}_{\text{orig}} = 0.48$) and destination ($\hat{\beta}_{\text{dest}} = 0.59$) and decrease strongly with distance ($\hat{\beta}_{\text{dist}} = -2.32$) [2, 24]. The Fay–Herriot estimate is a shrinkage composite of the direct estimate and the regression prediction:

$$\hat{F}_{ij}^{\text{FH}} = \gamma_{ij}\hat{F}_{ij}^{\text{IPW}} + (1 - \gamma_{ij})\mathbf{x}_{ij}'\hat{\boldsymbol{\beta}} \tag{4}$$

where the shrinkage weight $\gamma_{ij} = \hat{\sigma}_u^2/(\hat{\sigma}_u^2 + \psi_{ij})$ ranges from near 1 (trust the data, low sampling variance) to near 0 (trust the model, high sampling variance).

4 Results

We now summarise the main findings across the six pipeline steps.

Sector tessellation.

The sector-aware tessellation produces 5,378 sectors at 1,292 physical sites. The median effective resolution improves from 894 m (tower-point Voronoi) to 299 m (sector polygons), representing a

3.0× gain in spatial precision (Figure 2). This gain is driven by multi-sector sites: tri-sector and six-sector configurations together account for 53% of all sites. The 66 remaining omni-directional sites, by contrast, show no improvement over the Voronoi baseline.

Detection floor.

The detection floor varies by more than two orders of magnitude across R13 (Table 1). The distribution of d_{50} is heavily right-skewed, with a median of 63 m, a 95th percentile of 403 m, and a maximum of 2,542 m. The bulk of the network provides fine spatial resolution; the bias is concentrated in a relatively small number of peripheral sites serving large rural areas.

Intra-site mobility.

We find that intra-site transitions account for approximately 40% of all matched cell transitions (354 million of 890 million). After excluding technology handoffs (254 million transitions with no spatial content), the remaining 100 million sector crossings have a median estimated displacement of 429 m, partially filling the sub-kilometre gap in the displacement distribution that is entirely absent from inter-site-only analysis (Figure 5).

IPW correction.

The inverse probability weighting produces a clear spatial pattern of correction: rural comunas in the southwest and south of R13 receive 50–73% more trips after correction, while the urban core is modestly downweighted (∼0.86×) by the Hájek normalisation (Figure 6). This pattern closely tracks the spatial distribution of tower density, as expected.

Fay–Herriot smoothing.

The gravity model underlying the Fay–Herriot estimator explains the majority of the inter-comuna flow variance, with all four covariates significant at $p < 0.001$. Shrinkage is pronounced: the median γ across all OD pairs is 0.15, indicating that most pairs are dominated by the structural model rather than the direct IPW estimate. As anticipated, urban comunas retain high shrinkage weights ($\gamma \approx 0.7$, meaning the data are trusted), while peripheral comunas are pulled almost entirely toward the gravity prediction ($\gamma < 0.1$; Figure 7).

The three-way comparison in Table 3 illustrates the correction pipeline at work. For Santiago (urban), the IPW and Fay–Herriot estimates are close to the raw counts, because detection is near-complete and the data are reliable. For Pirque (rural), the IPW correction uplifts flows by 51%, and the Fay–Herriot model reinforces this adjustment slightly (+2%), reflecting the low shrinkage weight ($\gamma = 0.28$) that balances the noisy direct estimate against the structural model.

We note, however, that the variance reduction from the Fay–Herriot step is modest: only 29% of comunas show a lower coefficient of variation after smoothing. This suggests that the simple four-covariate gravity specification is adequate for large flows but introduces systematic deviations for the long tail of small inter-comuna pairs. A richer covariate set (incorporating road-network distance rather than Euclidean distance, administrative contiguity, and intervening opportunities) would likely improve the model for these cases.

5 Discussion

Relation to existing work.

The biases documented here are implicit in any analysis of tower-level mobile phone data, yet they are rarely quantified explicitly. Our contribution is not any single correction technique (census calibration, dasymetric interpolation, IPW, and Fay–Herriot estimation are all established methods [4, 6, 8, 13, 15, 23]) but rather their integration into a coherent pipeline that is parameterised by the physical properties of the tower network (antenna azimuth, height, and sectorisation) rather than by ad hoc tuning. The sector-aware tessellation, in particular, is a simple but effective use of catalog information that is available for most networks but seldom exploited [9].

Limitations.

Several limitations should be kept in mind when interpreting these results. First, we use data from a single operator; multi-operator coverage would allow estimation of market share rather than assuming it is uniform. Second, the study area is the most urbanised region in Chile, and the biases documented

here, while substantial, would likely be more severe in regions with lower tower density. Third, recent ground-truth travel-survey data for Santiago are limited. The last official origin-destination survey available for Santiago is the EOD 2012, whose results were released in 2015 [18], while the newer EMS 2024 is a useful mobility survey but not a full substitute for an origin-destination survey [14]; accordingly, our Monte Carlo simulation supports the internal consistency of the detection model but not its external accuracy against a contemporaneous benchmark. Fourth, the Fay–Herriot model employs a simple gravity specification that underperforms for small flows; richer covariates and spatial random effects would be natural extensions.

Generalisability.

The pipeline is designed to be transferable. It requires only three inputs that are available in virtually any country: a tower catalog with azimuth and height fields, a census or population layer, and a study area boundary polygon. The population layer should be fine enough that individual administrative units do not span many towers; if a single unit covers a large number of antennas, the uniform-density assumption within that unit weakens the IPW penetration-rate estimate. In practice, block- or neighbourhood-level data (as used here) are ideal, but even district-level layers are workable provided the study area is not extremely rural. The XDR transition data are needed only for the OD construction and correction steps (Steps 4–6); the spatial characterisation of the bias (Steps 1–3) can be carried out with the tower catalog and census alone. The accompanying `mobilens` library handles CRS detection automatically and exposes all column names and thresholds as configurable parameters.

Implications.

The 50–73% uplift in rural flows has direct consequences for downstream applications. In metapopulation models of disease spread, for example, uncorrected OD matrices underestimate the connectivity of rural communities and may lead to underestimation of epidemic arrival times and final attack rates outside the urban core [16, 17]. Similar concerns apply to transport planning, migration estimation [7], and any application that relies on accurate inter-regional flow magnitudes.

6 Conclusions

We have shown that the heterogeneous spatial distribution of cell towers in mobile phone networks creates a systematic, spatially structured measurement bias that underestimates rural mobility. Our six-step correction pipeline (sector tessellation, detection floor modelling, dasymetric population redistribution, intra-site mobility recovery, inverse probability weighting, and Fay–Herriot small area estimation) characterises and corrects this bias using standard inputs available in most countries.

Applied to the Santiago metropolitan region, the pipeline reveals that the 50% detection threshold varies from 16 m in the urban core to 2,542 m at the most isolated site; that 40% of observed cell transitions are intra-site, of which 100 million represent genuine sector crossings with a median displacement of 429 m; and that IPW correction uplifts rural comuna flows by 50–73%. These corrections are substantial enough to matter for any quantitative use of the data. The pipeline and all analysis code are available in the open-source `mobilens` Python library. The materials and source code to replicate this work are available at https://github.com/leoferres/towerdensity.

References

[1] Alexander L, Jiang S, Murga M, et al (2015) Origin–destination trips by purpose and time of day inferred from mobile phone data. Transportation Research Part C: Emerging Technologies 58:240–250. https://doi.org/10.1016/j.trc.2015.02.018

[2] Barbosa H, Barthelemy M, Ghoshal G, et al (2018) Human mobility: Models and applications. Physics Reports 734:1–74. https://doi.org/10.1016/j.physrep.2018.01.001

[3] Boeing G (2017) OSMnx: New methods for acquiring, constructing, analyzing, and visualizing complex street networks. Computers, Environment and Urban Systems 65:126–139. https://doi.org/10.1016/j.compenvurbsys.2017.05.004

[4] Briggs DJ, Gulliver J, Fecht D, et al (2007) Dasymetric modelling of small-area population distribution using land cover and light emissions data. Remote Sensing of Environment 108(4):451–466. https://doi.org/10.1016/j.rse.2006.11.020

[5] Calabrese F, Diao M, Di Lorenzo G, et al (2013) Understanding individual mobility patterns from urban sensing data: A mobile phone trace example. Transportation Research Part C: Emerging Technologies 26:301–313. https://doi.org/10.1016/j.trc.2012.09.009

[6] Deville JC, Särndal CE (1992) Calibration estimators in survey sampling. Journal of the American Statistical Association 87(418):376–382. https://doi.org/10.1080/01621459.1992.10475217

[7] Elejalde E, Ferres L, Navarro V, et al (2024) The social stratification of internal migration and daily mobility during the covid-19 pandemic. Scientific Reports 14(1). https://doi.org/10.1038/s41598-024-63098-5

[8] Fay RE, Herriot RA (1979) Estimates of income for small places: An application of James–Stein procedures to census data. Journal of the American Statistical Association 74(366a):269–277. https://doi.org/10.1080/01621459.1979.10482505

[9] Gallotti R, Maniscalco D, Barthélemy M, et al (2024) Distorted insights from human mobility data. Communications Physics 7:421. https://doi.org/10.1038/s42005-024-01909-x

[10] González MC, Hidalgo CA, Barabási AL (2008) Understanding individual human mobility patterns. Nature 453(7196):779–782. https://doi.org/10.1038/nature06958

[11] Graells-Garrido E, Peredo O, García J (2016) Sensing urban patterns with antenna mappings: The case of santiago, chile. Sensors 16(7):1098. https://doi.org/10.3390/s16071098

[12] Haraguchi M, Nishino A, Kodaka A, et al (2022) Human mobility data and analysis for urban resilience: A systematic review. Environment and Planning B: Urban Analytics and City Science 49(5):1507–1535. https://doi.org/10.1177/23998083221075634

[13] Horvitz DG, Thompson DJ (1952) A generalization of sampling without replacement from a finite universe. Journal of the American Statistical Association 47(260):663–685. https://doi.org/10.1080/01621459.1952.10483446

[14] Hurtubia R, Waintrub N, Raveau S (2024) Encuesta de movilidad de santiago 2024. https://www.cedeus.cl/blog/2024/08/21/encuesta-de-movilidad-de-santiago-2024/, cEDEUS report page, accessed: 2026-03-28

[15] Järv O, Tenkanen H, Toivonen T (2017) Enhancing spatial accuracy of mobile phone data using multi-temporal dasymetric interpolation. International Journal of Geographical Information Science 31(8):1630–1651. https://doi.org/10.1080/13658816.2017.1287369

[16] Jiang S, Yang Y, Gupta S, et al (2016) The TimeGeo modeling framework for urban mobility without travel surveys. Proceedings of the National Academy of Sciences 113(37):E5370–E5378. https://doi.org/10.1073/pnas.1524261113

[17] Kostandova N, Schluth C, Arambepola R, et al (2024) A systematic review of using population-level human mobility data to understand SARS-CoV-2 transmission. Nature Communications 15:1–12. https://doi.org/10.1038/s41467-024-54895-7

[18] Ministerio de Transportes y Telecomunicaciones de Chile (2015) Presentamos resultados de la encuesta origen destino de santiago. https://mtt.gob.cl/presentamos-resultados-de-la-encuesta-origen-destino-de-santiago/, accessed: 2026-03-28

[19] Naushirvanov T, Elejalde E, Kalimeri K, et al (2025) Evacuation patterns and socioeconomic stratification in the context of wildfires. EPJ Data Science 14(1). https://doi.org/10.1140/epjds/s13688-025-00540-2

[20] Pappalardo L, Ferres L, Sacasa M, et al (2021) Evaluation of home detection algorithms on mobile phone data using individual-level ground truth. EPJ Data Science 10(1). https://doi.org/10.1140/epjds/s13688-021-00284-9

[21] Pappalardo L, Cornacchia G, Navarro V, et al (2023) A dataset to assess mobility changes in Chile following local quarantines. Scientific Data 10:6. https://doi.org/10.1038/s41597-022-01893-3

[22] Pappalardo L, Manley E, Sekara V, et al (2023) Future directions in human mobility science. Nature Computational Science 3(7):588–600. https://doi.org/10.1038/s43588-023-00469-4

[23] Pfeffermann D (2013) New important developments in small area estimation. Statistical Science 28(1):40–68. https://doi.org/10.1214/12-STS395

[24] Simini F, González MC, Maritan A, et al (2012) A universal model for mobility and migration patterns. Nature 484(7392):96–100. https://doi.org/10.1038/nature10856

Software availability. The `mobilens` Python library (v0.1.0) implements the full pipeline. Source code is available at https://github.com/leoferres/mobilens under an MIT licence.

Declarations

Availability of data and materials. The tower catalog is proprietary and cannot be shared. The 2024 Chilean census cartography is publicly available from the Instituto Nacional de Estadísticas (INE). The OpenStreetMap road network is open data. The `mobilens` library and analysis notebooks are available at https://github.com/leoferres/mobilens.

Competing interests. The author declares no competing interests.

Funding. This research was supported by FONIS grant SA24I0124 to L.F. L.F. also acknowledges financial support from the Lagrange Project of the Institute for Scientific Interchange Foundation (ISI Foundation), funded by the Fondazione Cassa di Risparmio di Torino (Fondazione CRT).

Authors' contributions. LF and EE designed the study, developed the methodology, implemented the software, performed the analysis, and wrote the manuscript.

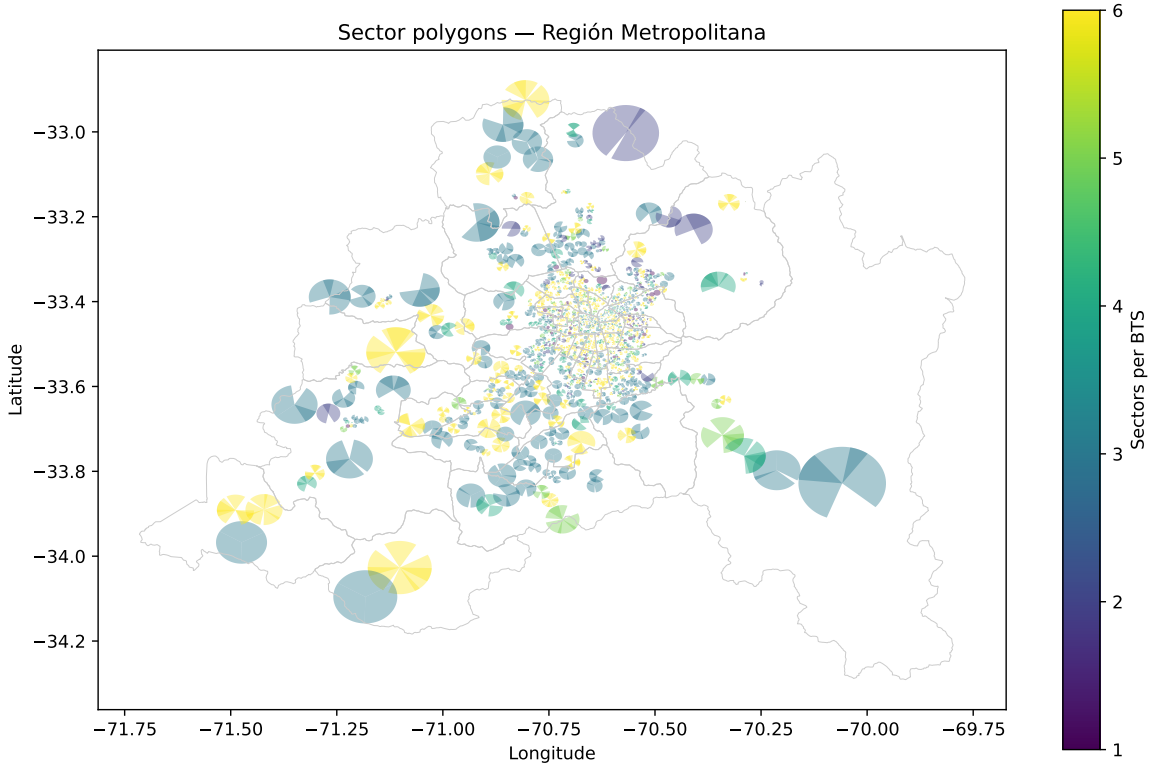

Fig. 1 Sector polygons across R13, coloured by the number of sectors per site. The contrast between the dense, small sectors in central Santiago and the sparse, large wedges at the periphery visualises the spatial heterogeneity that drives the tower-density bias. Each sector's radius is the minimum of the antenna's radio horizon (derived from mast height) and 65% of the nearest-neighbour tower distance; the angular width is $360°/n$, where n is the number of sectors at the site.

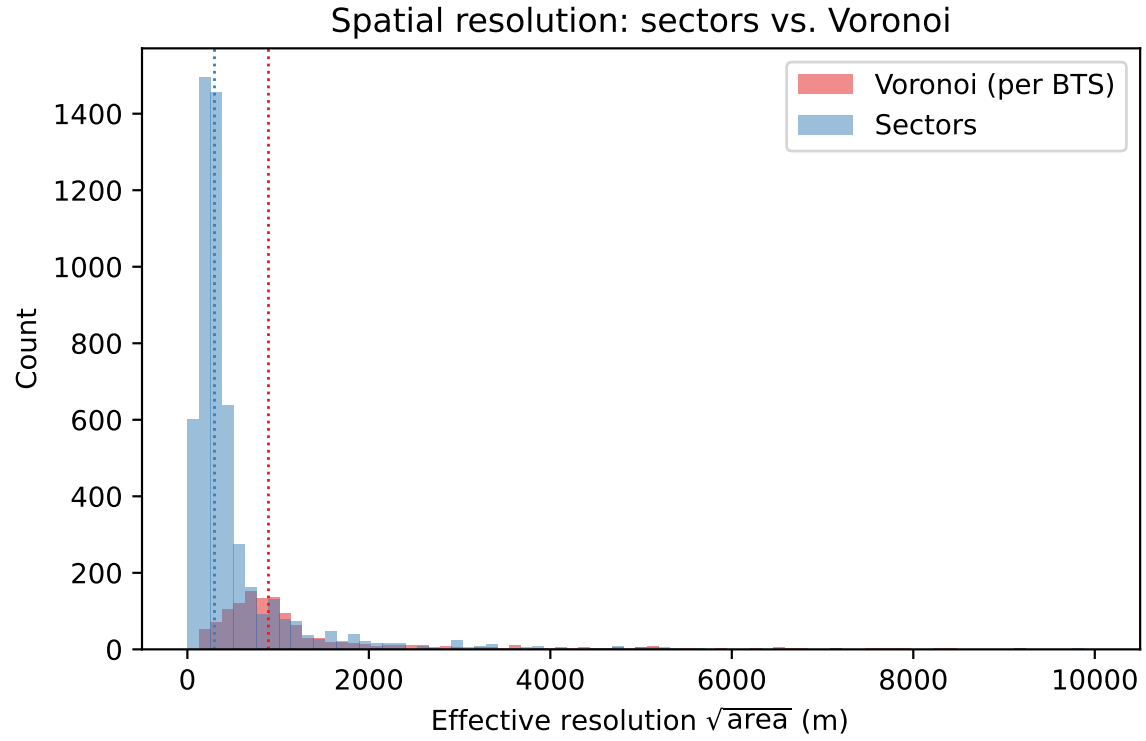

Fig. 2 Effective spatial resolution ($\sqrt{\text{area}}$) for sector polygons (blue) versus tower-point Voronoi cells (red). Dotted lines indicate medians: 299 m for sectors, 894 m for Voronoi, representing a 3.0× gain.

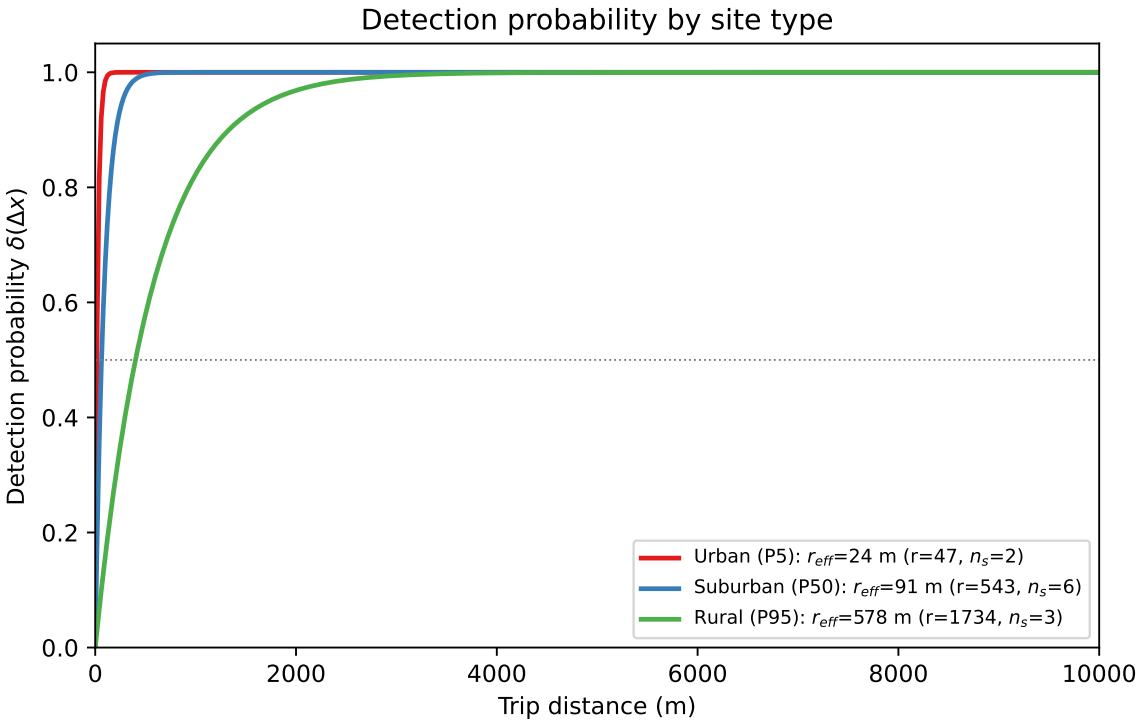

Fig. 3 Detection probability $\delta(\Delta x)$ as a function of trip distance for three example sites at the 5th, 50th, and 95th percentiles of effective resolution. Because tower density is highest in the urban core and lowest at the periphery, low percentiles of r_{eff} correspond to dense urban sites and high percentiles to sparse rural ones. The horizontal grey line marks the 50% detection threshold.

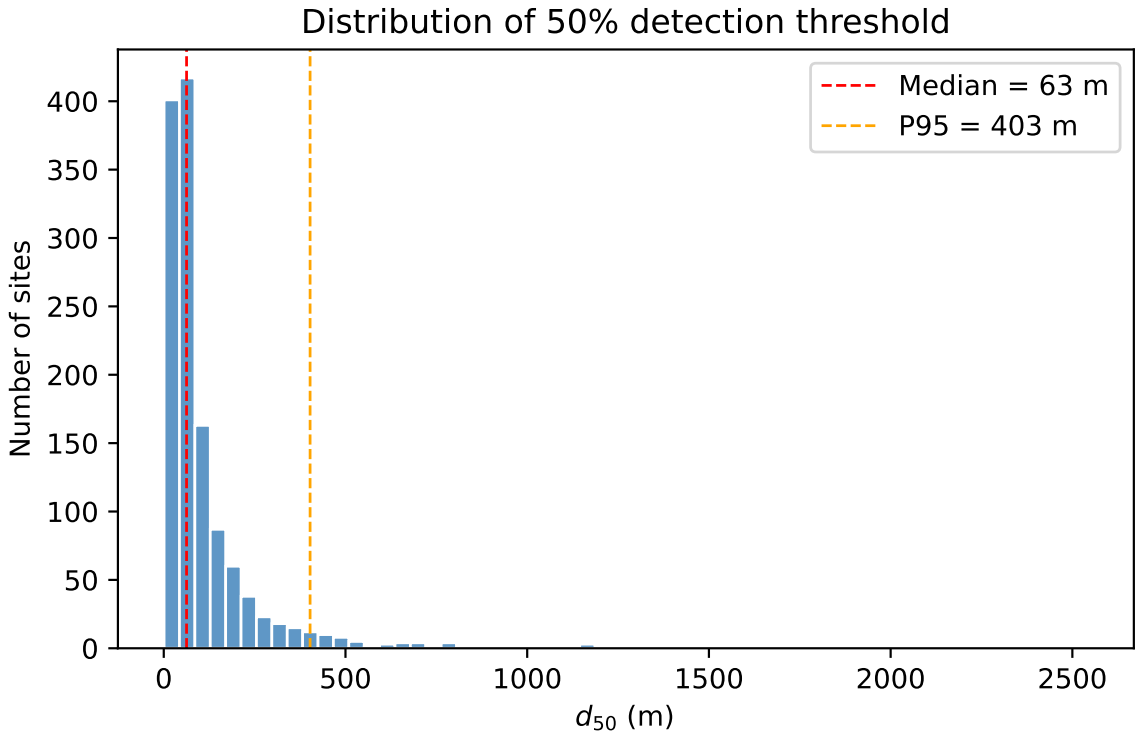

Fig. 4 Distribution of the 50% detection threshold d_{50} across 1,292 physical sites. The distribution is heavily right-skewed: median 63 m (red dashed), 95th percentile 403 m (orange dashed).

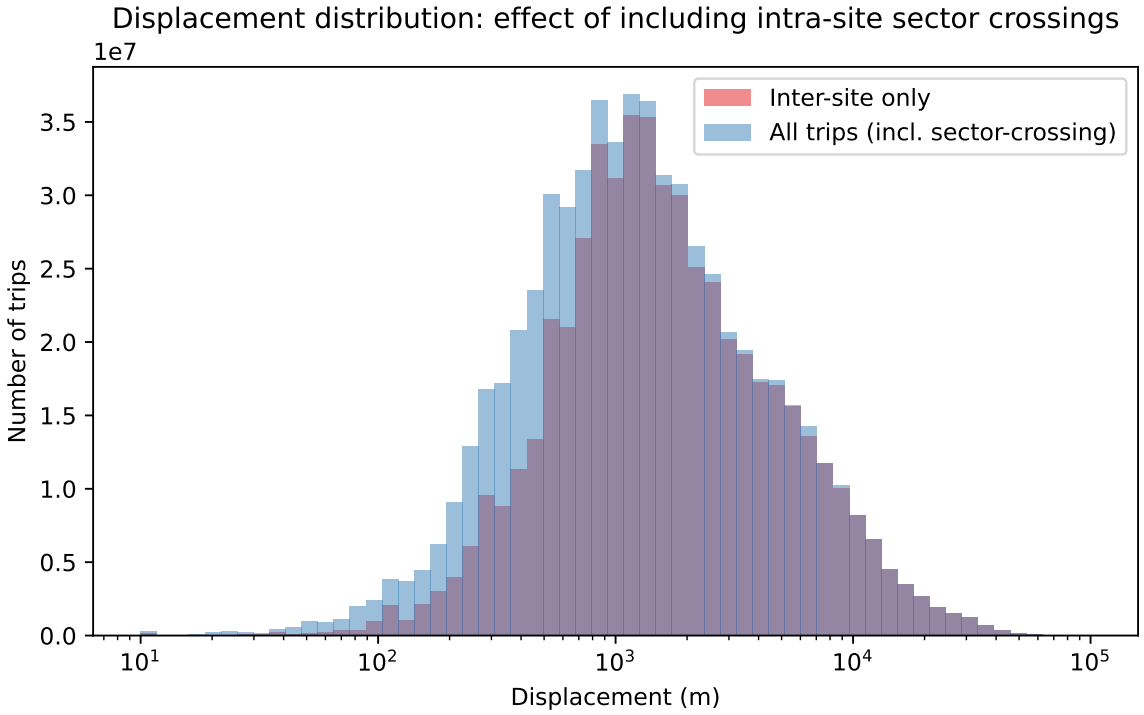

Fig. 5 Displacement distribution with (blue) and without (red) intra-site sector crossings. Including sector crossings fills the 100–500 m range that is absent from inter-site-only data.

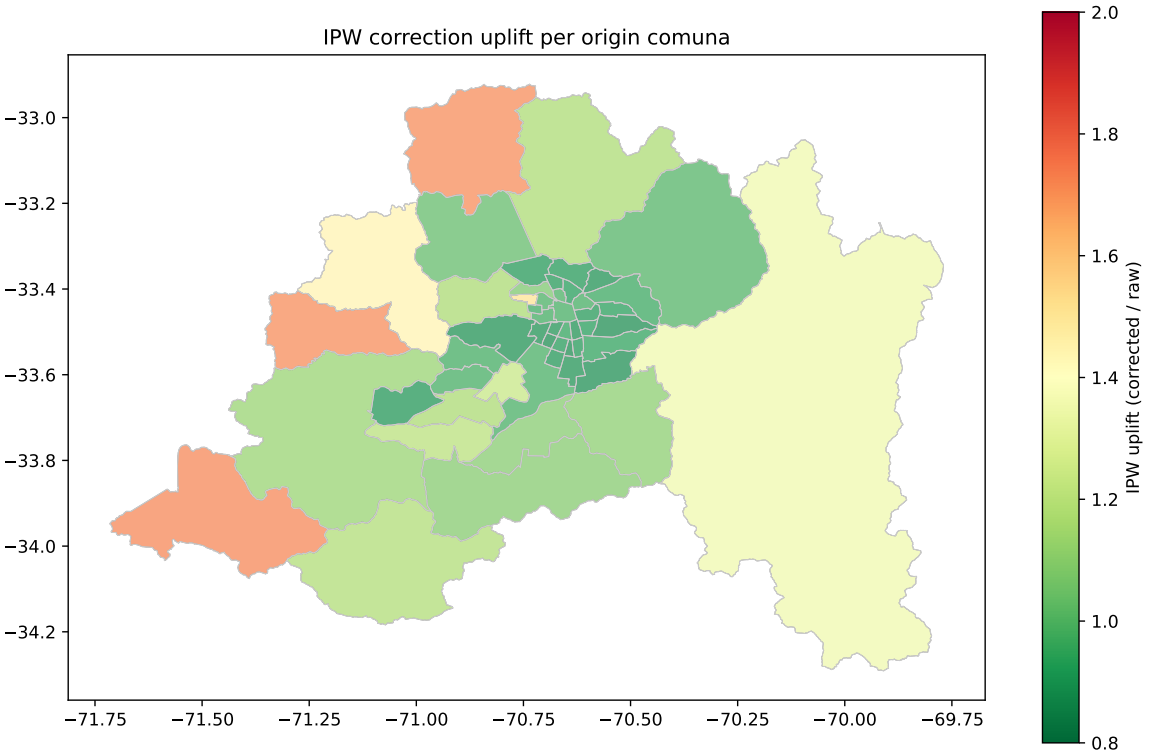

Fig. 6 IPW correction uplift (corrected / raw trip ratio) per origin comuna. Rural comunas in the southwest and south of R13 receive 50–73% more trips after correction; the urban core is slightly downweighted (∼0.86×).

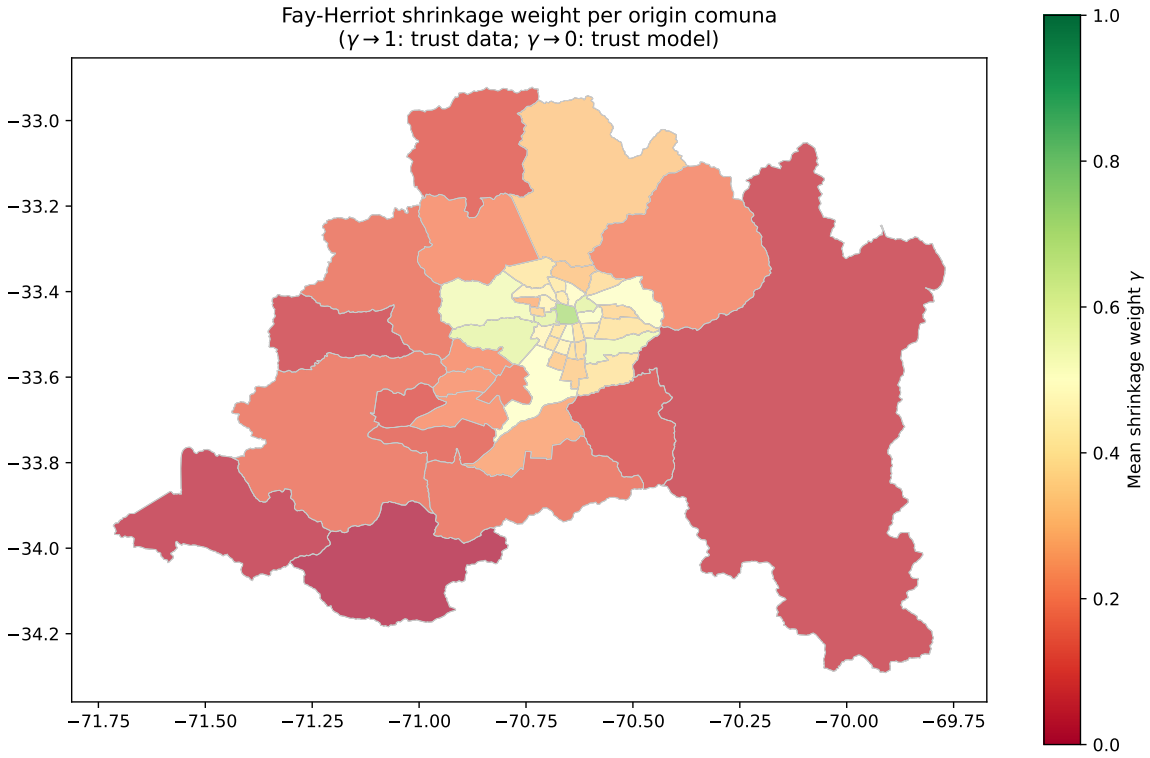

Fig. 7 Fay–Herriot shrinkage weight γ per origin comuna. Green ($\gamma \to 1$): the direct IPW estimate is trusted. Red ($\gamma \to 0$): the estimate is shrunk toward the gravity model prediction. The spatial gradient reflects the tower-density gradient.

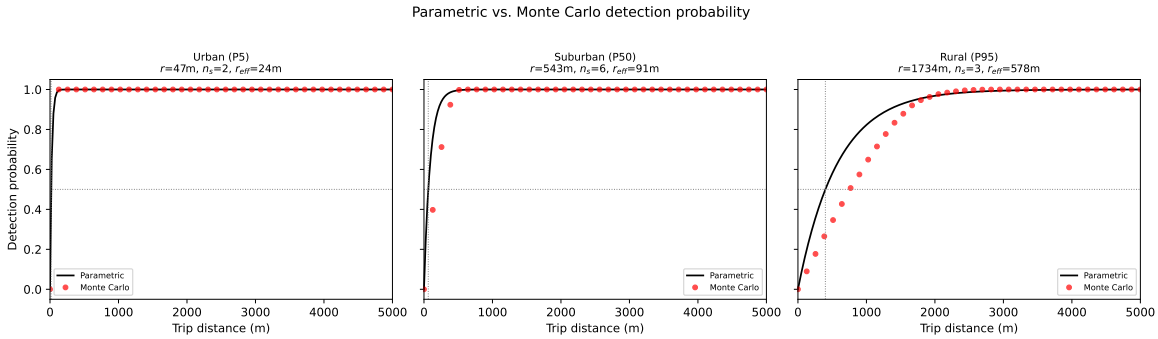

Fig. 8 Monte Carlo validation of the parametric detection probability formula for urban (left), suburban (centre), and rural (right) example sites. The black line is the parametric prediction; red dots are simulated detection rates. The parametric model is a conservative lower bound.